Genotype imputation in diverse populations: Empirical and theoretical approaches

by

Juichi (Lucy) Huang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2011

Doctoral Committee:

Associate Professor Noah A. Rosenberg, Co-Chair
Associate Professor Sebastian K. Zöllner, Co-Chair
Professor Michael L. Boehnke
Professor Stephen B. Gruber
Assistant Professor Jun Z. Li

To my parents,

my advisor,

&

my best friend and husband

ACKNOWLEDGEMENTS

"Happy news!" Dr. Mike Boehnke said over the phone. It was the first time I spoke with Mike, on a sunny afternoon in 2005. He informed me that I had been accepted as a Master's student in the biostatistics program at the University of Michigan, the first of many happy news I would hear on this fulfilling academic journey.

Shortly after I came to Michigan, I came to know Drs. Sebastian Zöllner and Noah Rosenberg from their population genetics class. I didn't know then Noah would later become not only my graduate advisor, but also a great friend. I have also been very fortunate to befriend many of the students, staff, and faculty from both biostatistics and bioinformatics programs. With their generous help and support, I have matured as an individual and as a researcher.

This dissertation would have been next to impossible without contributions of many individuals. My deepest gratitude goes to Noah, an outstanding researcher and an amazing mentor. His passion for learning and teaching is contagious, and his attention to detail is impressive. Day after day, he has shown me—both in words and in actions—how to conduct research with integrity, rigor, and originality. To Noah, science and life are inseparable: there is a mathematical explanation for every empirical observation. He draws inspirations from all aspects of life, and he naturally inspires those around him too to be intellectually curious about the world around them. Noah is also incredibly generous with his ideas and time. He has gone above and beyond his duty as an advisor to help me succeed in everything I do. Without Noah's patience, faith, and support, it would have been difficult, if not impossible,

for me to complete the doctoral training. Over the years, Noah has supported me wholeheartedly through the ups and downs of my academic and personal lives. Noah is and will always be one of my role models. I am grateful for his dedication and commitment to students and hope to make him a proud teacher in all my future endeavors.

I would also like to thank my dissertation committee co-chair, Dr. Sebastian Zöllner, and the other members of the committee, Drs. Mike Boehnke, Stephen Gruber, and Jun Li, for their helpful comments and insightful questions. I would like to extend a special thanks to Sebastian and Mike for their warm support throughout my entire graduate career.

I would like to thank all members of the Rosenberg lab for making the lab feel like a second home. A very special thanks goes to Trevor Pemberton for preparing many of the datasets used in this dissertation, and for baking goodies all year round. Thank you, Trevor, for officiating my wedding twice, with a small group of friends in Ann Arbor and with much larger extended families and friends in Hawaii. I want to thank Mike DeGiorgio, Zach Szpiech, and Ethan Jewett for their constant encouragement and for countless discussions of research ideas and progress in our cubicles. Thank you, Mike and Ethan, especially for reading parts of this dissertation, and for providing feedback and constructive criticisms, and thank you, Zach, for showing me example codes of multi-threaded programming. I would also like to extend my appreciation to my lab members' significant others, in particular, Ekjyot Saini, Raquel Assis, and Marie-France Mifune, for their friendship and for a wonderful surprise baby shower.

I would like to thank the individuals who have guided and contributed to my research. First, I would like to thank Dr. Vanja Dukic for her mentorship in my undergraduate research and for solidifying my research interests in health sciences. I would like to acknowledge Drs. Margit Burmeister and Dan Burns for their helpful conversations and assistance with my transfer to the bioinformatics program. I

would also like to thank the IT administrators at CCMB for working hard to ensure seamless connectivity and availability of the computing environment, as well as accommodating many of the computing requests from the Rosenberg lab. Finally, I would like to acknowledge the co-authors of Chapters II-V of this dissertation for their invaluable contributions: Chapter II—Drs. Yun Li, Andrew Singleton, John Hardy, Gonçalo Abecasis, Noah Rosenberg, and Paul Scheet; Chapter III—Chaolong Wang and Dr. Noah Rosenberg; Chapter IV—Drs. Mattias Jakobsson, Trevor Pemberton, Muntaser Ibrahim, Thomas Nyambo, Sabah Omar, Jonathan Pritchard, Sarah Tishkoff, and Noah Rosenberg; Chapter V—Drs. Erkan Buzbas and Noah Rosenberg.

I want to thank my parents, Ken Huang and Helen Hsieh, for their unconditional love and support. My father, who places a strong emphasis on education, has motivated me to strive for academic excellence and to never stop learning. My mother, both a great listener and friend, has encouraged me to treat every moment as my last and to live life to its fullest. As life-time entrepreneurs, they have instilled in me a strong sense of optimism and taught me to face every challenge in life with a smile.

Finally, I want to thank my best friend and husband, Jack Chen, for everything he has done for me. A former software engineer, Jack has provided on-demand refresher courses on many programming languages, as well as an impeccable IT service and live assistance with computing issues at home. With keen observations and insights, Jack has shown and encouraged me to employ out-of-box approaches to problem solving. Most importantly, his love, generosity, and sincerity have inspired me to be a better person everyday. Jack is a wonderful person, husband, and father, and I am grateful to have him and his families that deeply care for us in my life. I thank our son, Aiden, too for delaying his entrance to this world to allow the completion of this dissertation, and for showering Jack and me with so much joy since his birth.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

The development of high-throughput genotyping technologies has revolutionized the process of identifying disease-predisposing genetic variants. In particular, the vast quantities of genotype and DNA sequence data generated by these technologies have enabled geneticists to closely examine statistical correlations between genetic markers and common complex traits via genome-wide association (GWA) studies. Since their inception in 2005 (e.g., Klein *et al.*, 2005), GWA studies have identified over 4,900 single-nucleotide polymorphism (SNP) loci associated with common diseases and other complex traits (Hindorff *et al.*, 2009, 2011). Some of these association signals have led to the discovery of previously unsuspected etiological pathways for common diseases, thereby uncovering potential molecular targets for therapeutic applications (e.g., Montes *et al.*, 2009; Klionsky, 2009; Yano & Kurata, 2009). However, most currently known genes that underlie the predisposition to complex diseases were discovered primarily through studies of European populations, which contain only a subset of human genetic variation. Worldwide differences in frequencies of disease alleles, effect sizes of risk variants, and occurrences of rare variants can affect the detectability and importance of risk-modifying genes in different populations (Rosenberg *et al.*, 2010). Thus, questions have arisen about the generalizability of existing GWA findings across populations of diverse ancestry.

Expanding the search for genes that influence human diseases from European to non-European populations addresses the extent to which current GWA discoveries are generalizable to worldwide human populations (Need & Goldstein, 2009; Rosenberg *et al.*, 2010; Bustamante *et al.*, 2011). Such an expansion provides a mechanism to prioritize the discoveries for an evaluation of their diagnostic and prognostic potential, and helps to identify promising genetic variants that affect multiple populations of different ancestry. Additionally, risk variants may have different effect sizes or allele frequencies across populations, suggesting that different determinants of the same disease may exist in different populations (Tang, 2006; Adeyemo & Rotimi, 2010). In the case in which GWA discoveries are unlikely to be generalizable across populations, the search for risk-promoting genes specific to particular populations is warranted, in order to contribute toward alleviating the disease burden in those populations. Finally, as large-scale genetic studies begin to produce clinically actionable results, it is important to ensure that populations of diverse ancestry are included in such studies to avoid exacerbating health-care disparities (Need & Goldstein, 2009).

A recent advance—genotype imputation (Nicolae, 2006; Li *et al.*, 2006; Browning & Browning, 2007; Marchini *et al.*, 2007; Servin & Stephens, 2007)—holds one of the keys for expanding the collection of disease-association signals initially detected in European populations. Genotype imputation is a statistical approach that leverages high-resolution genetic data from reference datasets to predict genetic variants at marker positions not directly measured in a particular GWA study. This prediction, or imputation, both increases the number of markers that can be directly tested for disease associations and enables the integration of datasets from multiple GWA studies. Thus, imputation has the potential to increase the statistical power to detect disease-susceptibility genes and to facilitate statistically robust identification of these genes through larger sample sizes.

Crucial to the success of imputation procedures is the representation of the popula-

tion under consideration in reference datasets that contain "template" sequences from which missing genotypes in GWA samples are inferred (Egyud *et al.*, 2009; Huang *et al.*, 2009a; Paşaniuc *et al.*, 2010; Shriner *et al.*, 2010). However, large-scale genomic datasets, such as those from the International HapMap Project (2005; 2007; 2010) and the 1,000 Genomes Project (2010), only exist for a limited number of populations. One way of aiding the search for complex-disease genes in many non-European populations for which reference datasets do not exist is to employ innovative imputation strategies based on statistical and population-genetic principles.

This dissertation centers on developing genotype-imputation strategies that optimize the use of existing genomic resources for genetic studies of complex diseases in diverse human populations. In Chapters II (Huang *et al.*, 2009a) and III (Huang *et al.*, 2009b), I pursue the following objectives: (1) I evaluate the portability of existing genomic resources for imputation-based GWA studies in 29 worldwide populations; (2) I identify the optimal combinations or mixtures of existing reference datasets for enhancing imputation in these diverse populations; and (3) I quantify the increase in the minimal sample size, due to imperfect imputation, that would be required to provide the same level of statistical evidence of disease predisposition for genetic variants that are imputed, rather than directly measured, in the 29 populations.

To achieve these objectives, I use genotype data on over 500,000 genetic markers in 443 individuals from 29 diverse human populations, as well as data on nearly two million genetic markers in 210 individuals from three reference populations. Through a series of imputation experiments, I find that African populations are generally the most difficult to impute accurately. Moreover, I find that nearly all populations benefit from the use of a mixture approach that I develop for selecting appropriate panels of reference data. Considering a simple 2×3 chi-squared test and the distribution of its non-centrality parameter, I next estimate the adjustments in the minimal sample size required to detect disease associations, at imputed versus directly measured

markers. Surprisingly, I find that even a 1% increase in imputation error can lead to a substantial increase (5-13%) in the sample size required for detecting risk-enhancing SNPs that are imputed. These results imply that advanced statistical and computational approaches that decrease imputation error can substantially reduce the sample sizes needed for imputation-based detection of genetic variants that underlie complex human diseases.

In agreement with observations from other studies of genotype imputation in globally distributed human populations (Guan & Stephens, 2008; Pei *et al.*, 2008; Huang *et al.*, 2009a, 2012; Li *et al.*, 2009; Fridley *et al.*, 2010; Surakka *et al.*, 2010), Chapters II and III also demonstrate that imputation accuracy varies across populations and that population-genetic factors play a role in determining the level of imputation accuracy attainable in a particular population. However, it has been unclear how these factors, such as the overall level of linkage disequilibrium in a population targeted for imputation and the degree of genetic similarity between the target and candidate reference populations, influence imputation accuracy. In Chapters IV (Huang *et al.*, 2012) and V, using empirical data and coalescent theory, respectively, I provide detailed investigations of the relationship between population-genetic factors and imputation accuracy.

In Chapter IV, using genotype data in 253 individuals from 15 Sub-Saharan African populations and in 901 individuals from 11 reference populations, I study how genotypic and haplotypic variation in population pairs containing a target population and a reference population affect imputation accuracy in the target population. I find that simple summary statistics of population differentiation, such as F_{st} between target and reference populations, correlate well with imputation accuracy. Thus, I recommend their use for predicting the optimal reference panel among a collection of candidate panels for imputation in the target. Furthermore, extending the investigation to an additional 854 individuals from 48 populations worldwide, I observe a

pattern in imputation accuracy that is consistent with models of out-of-Africa migrations of modern humans. Under these models, haplotype diversity is predicted to be greatest in Africa, potentially explaining the observation that imputation accuracy is higher when imputing untyped markers in a non-African population on the basis of an African population than when performing imputation in the reverse direction. Encouraged by the potential of population-genetic models to clarify the determinants of imputation accuracy, I next develop a population-genetic modeling framework to study genotype imputation.

In Chapter V, using a two-population demographic model in which a pair of populations diverged at some time in the past, I derive the approximate expectation and variance of imputation accuracy in a target sequence sampled from one of the two populations, using reference sequences sampled either from the same population as the target sequence or from the other population. I analytically show that under this model, imputation accuracy—as measured by the proportion of polymorphic sites that are imputed correctly in the target sequence—increases in expectation with the mutation rate, with the proportion of the markers in a chromosomal region that are genotyped in the target, and with the time to divergence between the target and reference populations. Each of these effects is likely to derive from an increase in information available for determining the reference sequence that is genetically most similar to the sequence targeted for imputation. I further analyze as a function of the divergence time the expected gain in imputation accuracy in the target using a reference sequence from the same population as the target rather than from the other population. I find that this expected gain in accuracy can be approximated with a simple expression consisting of quantities that describe the underlying genealogical relationship between target and reference sequences.

The research presented in this dissertation focuses on the development and application of genotype-imputation strategies for association studies of complex diseases

in diverse human populations. The first two published chapters are among the first attempts to offer recommendations on the design of reference panels for imputation-based genomic studies that seek to identify disease-predisposing genes in worldwide human populations. These study-design recommendations help address an important current issue in human genetics: whether disease findings obtained from European populations are generalizable to populations around the globe (Need & Goldstein, 2009; Rosenberg *et al.*, 2010; Bustamante *et al.*, 2011). The latter two chapters explore the empirical and theoretical basis for the ways in which population-genetic factors influence imputation accuracy. The results from these chapters provide insights on ways to improve forthcoming GWA analyses in some of the most genetically diverse human populations, such as those from Africa. Overall, this dissertation will facilitate future association studies in populations of diverse geographical origins, contributing to the mapping of genetic determinants of complex diseases for the human species as a whole.

CHAPTER II

Genotype Imputation Accuracy across Worldwide Human Populations

2.1 Introduction

The recent availability of high-density single-nucleotide polymorphism (SNP) genotype databases from several human populations has facilitated the mapping of complex disease loci in genome-wide association (GWA) studies. These databases, such as The International HapMap Project (2.5 to 4 million SNPs genome-wide; 2005; 2007) and SeattleSNPs ($\sim$7 Mb of resequencing data in genes), provide high-resolution information about allele frequencies and patterns of linkage disequilibrium (LD) among SNPs typed in the samples. They serve as "reference panels" useful for diverse purposes in human genetics.

Information in reference panels can be leveraged in a mapping context by merging the reference genotype data with collections of data from individual GWA studies (Figure 2.1). Because typical GWA studies contain genotype data on, at most, a few hundred thousand to a million SNPs, a very specific missing data pattern emerges from the union of a reference panel with a GWA data set. That is, for most SNPs, observations exist for the reference panel but not for the GWA study (Figure 2.1d). By modelling the pattern of LD in the reference panel, and then applying the fitted

model to the observed GWA study data, the "missing" GWA SNP genotypes can be effectively imputed (Li *et al.*, 2006; Nicolae, 2006; Marchini *et al.*, 2007; Servin & Stephens, 2007; Yu & Schaid, 2007; Browning, 2008; Guan & Stephens, 2008). Imputed genotypes at these SNP loci can then be used to test for association with disease in the same way that testing occurs for SNPs that were actually genotyped in the GWA study.

To date, most GWA studies have been conducted in populations that are well-represented by the available high-density reference panels. Specifically, study samples have typically derived from populations of northern European ancestry, for which the HapMap CEU panel – based on individuals of northern and western European descent sampled in Utah – has provided additional information for imputation in association testing (Scott *et al.*, 2007; Wellcome Trust Case Control Consortium, 2007; Reiner *et al.*, 2008; Willer *et al.*, 2008). However, for the purpose of genotype imputation, it is unclear how well the HapMap panels represent the patterns of genetic variation in other populations, particularly those that are more distant from the available panels, either in terms of demographic history or in terms of geographic proximity. Here we attempt to evaluate the "portability" of these panels for imputation-based studies of diverse human populations; this work is analogous to recent assessments of the portability of informative SNPs chosen from reference panels in providing LD-based genomic coverage in diverse populations (Conrad *et al.*, 2006; González-Neira *et al.*, 2006; Gu *et al.*, 2007, 2008; Xing *et al.*, 2008).

Recently two studies examined patterns of SNP variation in multiple human populations from around the world, providing data on samples from the Human Genome Diversity Project (HGDP) at more than 500,000 SNPs (Jakobsson *et al.*, 2008; Li *et al.*, 2008). We select one of these databases (Jakobsson *et al.*, 2008) and in several ways we evaluate the behavior of a missing data imputation algorithm in each of the sampled populations. First, using the sampled populations alone, we assess

average imputation accuracy when imputing masked genotypes in the absence of a reference panel (Figure 2.1a). Second, we use the European American (CEU), Yoruba (YRI), and combined Chinese and Japanese (CHB+JPT) panels from the HapMap project in various combinations as reference panels, and we evaluate the properties of imputation in the sampled populations using the reference panel data (Figures 2.1b and 2.1c). Finally, using data from a targeted high-density scan of several genomic regions on chromosome 21 in the HGDP samples (Conrad *et al.*, 2006; Pemberton *et al.*, 2008), we also assess the accuracy with which genotypes of untyped markers can be imputed in these populations from the ∼500,000 typed SNPs and various combinations of HapMap reference panels (Figure 2.1d).

We find that when employing HapMap reference panels for imputation, genotypes from European HGDP samples are imputed with the highest accuracy, followed by samples from East Asia, Central/South Asia, the Americas, Oceania, the Middle East, and Africa. The choice of preferred HapMap reference panels for imputation in populations worldwide follows major geographic groupings. For most HGDP populations, we obtain additional gains in imputation accuracy when imputing genotypes based on a mixture of available reference panels. These findings can serve as a basis for the application of imputation methods to analysis of genomic data in populations worldwide.

2.2 Materials and Methods

2.2.1 Data

We examined a subset of 443 unrelated individuals from 29 populations in the HGDP-CEPH Human Genome Diversity Cell Line Panel, a worldwide collection of individuals from diverse locations (Cann *et al.*, 2002). Individual genotypes obtained using the Illumina HumanHap550 SNP platform had been previously reported by Jakobs-

son *et al.* (2008) at 513,008 biallelic autosomal genetic markers (246 SNPs ultimately discarded by Jakobsson *et al.* (2008) to produce their final data set of 512,762 SNPs were included here as potentially informative for imputation).

For some analyses, we incorporated additional individuals to serve as reference data for imputing missing genotypes. The reference data consisted of phased haplotypes of 210 individuals from the International HapMap Project (2005; 2007): 60 European Americans sampled from Utah, USA (abbreviated CEU), 60 Yoruba individuals from Ibadan, Nigeria (YRI), 45 Chinese from Beijing, China, and 45 Japanese from Tokyo, Japan (CHB+JPT). The phased HapMap data (release 21) were downloaded from the HapMap phase II data website (see Web Resources). The CHB and JPT haplotypes were combined into a single panel, and the specific origins of individual haplotypes (either CHB or JPT) were ignored. The CEU and YRI sets consisted of parents from trios; the offspring were omitted from our study but had been used in inferring haplotypes in the parents. A total of 1,958,375 autosomal markers polymorphic in the set of 210 HapMap individuals were used in our analyses. All except two of these SNPs (rs7008731 and rs13332778) were separately polymorphic in the CEU, YRI, and CHB+JPT panels.

In some analyses, we used data from Conrad *et al.* (2006), where some of the genotypes imputed with the data of Jakobsson *et al.* (2008) were measured directly in the same HGDP samples. These analyses used an updated version of the Conrad *et al.* (2006) data from Pemberton *et al.* (2008).

2.2.2 LD-based Imputation

Multiple models exist for accurate imputation of missing genotypes based on LD information (Nicolae, 2006; Marchini *et al.*, 2007; Servin & Stephens, 2007; Browning, 2008; Stephens & Scheet, 2005; Scheet & Stephens, 2006; Browning & Browning, 2007; Lin *et al.*, 2008). For our investigations of variation in genotype imputation

accuracy across populations, we used a recent implementation of a model related to the approach of Li & Stephens (2003), namely the Markov Chain Haplotyping algorithm (MACH-1.0.15) of Li *et al.* (2006); see Web Resources.

The intuition underlying this imputation approach is that collections of individuals, even those who are "unrelated," share short stretches of DNA sequence derived identically by descent from their common ancestors. Once these stretches are identified using a set of SNPs, it is possible to probabilistically predict alleles for intervening SNPs that are not measured in a given individual but that are measured in other individuals. Using a hidden Markov model, the algorithm resolves a collection of unphased genotypes into imperfect mosaics of several "template" haplotypes, from which it obtains an imputation or a best guess of each unknown genotype in each individual under consideration. All of our analyses rely on these "best guess" imputations, ignoring uncertainty in the genotype estimates. Exact software settings are given in the appendix.

2.2.3 Inferring Missing Genotypes without Additional Reference Individuals

To assess the impact of the proportion of missing genotypes on imputation accuracy in each population, we masked a fraction of the genotypes at random, and we then compared the estimated genotypes to the actual, masked genotypes (Figure 2.1a). The proportion of missing genotypes was varied between 5% and 50% with a 2.5% increment. That is, each diploid genotype was masked independently with probability equal to the specified proportion of missing genotypes. The proportion of correctly imputed *alleles* is reported as "imputation accuracy" throughout our analyses. For example, if the correct genotype was homozygous at a locus for a particular individual and a heterozygous genotype was imputed, then the algorithm was viewed as having produced 1 of 2 correct alleles. Similarly, if the algorithm imputed a homozygous

genotype at a locus where the correct genotype was heterozygous, then we considered the algorithm to have produced 1 of 2 correct alleles. It follows that the maximal number of alleles possible to impute incorrectly was 2 when the unknown genotype was homozygous and 1 when the unknown genotype was heterozygous.

In each of the 29 population samples, we measured the imputation accuracy for each proportion of missing genotypes, averaging across all markers. We summarized imputation accuracy genome-wide by the weighted average of chromosome-specific imputation accuracy, using the numbers of SNPs on individual chromosomes as the weights. In our analysis of the role of the proportion of missing genotypes, an individual's missing genotypes were estimated based on information strictly from other individuals in the same population sample. To obtain comparable results across populations, we restricted our analyses to a sample size of six individuals per population, the smallest sample size among the 29 populations. For each population, the six individuals were chosen randomly.

To evaluate the effect of sample size on imputation accuracy, we generated sub-samples for each population and each sample size by sequentially removing individuals one at a time from the full sample. To ensure that random sub-samples of individuals were used in the evaluation of imputation accuracy in each population, each of the population samples was permuted prior to the construction of sub-samples. In each data set, genotypes were hidden with a proportion of missing genotypes equal to 15%, and missing genotypes were estimated by MACH. We assessed imputation accuracy for various sample sizes for each population, and again summarized it by the weighted average allelic imputation accuracy across autosomes. As imputation accuracy varies across individuals in a population, the sequence in which individuals were removed from a full population sample could conceivably influence the relationship between imputation accuracy and sample size. Therefore, to examine the importance of the particular sequence of individuals utilized in the estimation procedure, we repeated

the analysis using a second randomly chosen sequence of individuals in each population. Differences in imputation accuracy from the two sequences (i.e., imputation accuracies based on the first permuted sample minus corresponding values based on the second permuted sample) were negligible for most populations and sample sizes (Figures S2.1 and S2.2).

2.2.4 Inferring Missing Genotypes with Additional Reference Individuals

2.2.4.1 Imputation Accuracy versus Panel Size

Using a single HapMap panel (either the CEU, YRI, or CHB+JPT sample) as a reference group to infer missing genotypes (Figure 2.1b), we investigated the relationship between imputation accuracy and reference panel size. For each HapMap panel, we permuted the panel and constructed random sub-panels of size 10, 20, ..., 120 haplotypes by sequentially adding 10 haplotypes in the order specified by our permutation. Note that each of the resulting sub-panels, when viewed independently, represented a random sample of haplotypes from the appropriate HapMap panel, and a consecutive pair of haplotypes did not necessarily correspond to two haplotypes of the same individual. To obtain comparable results across HapMap panels, we considered (only in this analysis) sub-panels of size at most 120 haplotypes, despite the fact that the CHB+JPT panel had size 180 haplotypes. In all populations, we utilized the same set of sub-panels derived from the HapMap samples for imputation. Based on each reference panel and its sub-panels, we performed genotype imputation and evaluated the accuracy across various sizes for a given reference panel as well as across reference panels for a given size. This analysis used the full sample from each HGDP population.

2.2.4.2 Imputation Accuracy versus Panel Composition

In addition to assessing imputation accuracy using each of the three HapMap panels in isolation, we also considered the panels combined together, and we considered other mixtures of the various panels (Figure 2.1c). To identify the mixture that produced the maximal imputation accuracy, we imputed missing genotypes in each population using mixed reference samples formed by combining individuals from the three HapMap groups. In contrast to our previous analyses, in which we considered missing genotypes on the entire autosomal genome, in this analysis we imputed only unknown genotypes on one chromosome, chromosome 2, in the interest of reducing computation time. We considered a variety of mixtures, with each mixture consisting of combinations of HapMap reference haplotypes chosen according to a specified ratio.

For each ratio, we used a reference panel of maximal size, constrained by the fact that most ratios involving two or more reference panels do not permit use of all available haplotypes from the panels under consideration. The set of mixtures that we considered corresponded to the set of vectors (i_1, i_2, i_3) of nonnegative integers with $i_1 + i_2 + i_3 = 7$. For each vector, we constructed a mixture sample consisting of a_1 CHB+JPT haplotypes, a_2 CEU haplotypes, and a_3 YRI haplotypes, so that a_1, a_2, and a_3 were as large as possible and so that they satisfied $a_1 : a_2 : a_3 = i_1 : i_2 : i_3$. For example, the vector $(i_1, i_2, i_3) = (4, 2, 1)$ led to $(a_1, a_2, a_3) = (180, 90, 45)$.

In each population, using all individuals sampled from the population, we assessed imputation accuracy using each of 36 mixed collections of haplotypes from the three HapMap panels (corresponding to the 36 solutions to $i_1 + i_2 + i_3 = 7$). For each (i_1, i_2, i_3), within HapMap groups, haplotypes were chosen randomly among the haplotypes present, and the same randomly chosen subsets of the three HapMap panels were used as the reference panel in all HGDP populations. The random sets of haplotypes were chosen so that if h haplotypes from a HapMap population were used in one mixed collection and $h' > h$ haplotypes from the same HapMap population were

used in another mixed collection, then it was always true that the set of h haplotypes comprised a subset of the set of h' haplotypes. For (i_1, i_2, i_3) given, the solution for the number of haplotypes, (a_1, a_2, a_3), was obtained as described in the appendix.

2.2.5 Application to Untyped Markers

In current GWA studies, genotypes are collected at densities on the order of ∼500,000 SNPs genome-wide. In such a study, by using a reference panel, additional information can be obtained about the genotypes of SNPs not typed directly in the GWA study but measured in an external reference panel. To assess the accuracy with which the genotypes of these markers can be imputed, we used the 513,008 SNPs typed in samples from 29 populations (Jakobsson *et al.*, 2008) in combination with the HapMap reference panels to impute genotypes of 1,445,367 SNPs. We then compared the imputed genotypes to those measured directly by Conrad *et al.* (2006) and updated by Pemberton *et al.* (2008), which, for limited regions of the genome, consist of SNPs at higher density than in a typical GWA study. Using this protocol, we assessed imputation accuracy at 218,345 diploid genotypes, as described below. We note that in contrast with our other analyses, in which genotypes were imputed in randomly chosen SNP positions that varied across individuals, in this analysis, for certain markers genotyped only in the reference panel, the genotypes of *all* individuals in the study sample were imputed. To distinguish this scenario from the "missing genotypes" scenarios of our other analyses, we refer to such markers as "untyped markers."

Among the 2810 SNPs reported by Pemberton *et al.* (2008), 1272 were located on chromosome 21, so we restricted this analysis to chromosome 21 for convenience. Among these 1272 SNPs, 1008 had not been included in the SNP set studied by Jakobsson *et al.* (2008) Of the 1008 SNPs, 513 were genotyped in the HapMap individuals. We thus assessed imputation accuracy at these 513 SNPs by using the genotypes at

6068 SNPs from Jakobsson *et al.* (2008) and the 26,716 SNPs available on chromosome 21 in the HapMap data. Using the HapMap reference panels to impute genotypes of untyped markers in all 443 individuals studied by Jakobsson *et al.* (2008), we measured imputation accuracy for the 513 SNPs in a set of 426 individuals. This set of 426 individuals is the intersection of the set of 927 unrelated HGDP individuals studied by Conrad *et al.* (2006) and Pemberton *et al.* (2008) with the set of 443 unrelated HGDP individuals studied by Jakobsson *et al.* (2008). The set contains at least five individuals from each of 29 populations. In total, of the 2(426)(513)=437,076 possible alleles in which imputation accuracy could be measured, 436,690 alleles were available (that is, 386 alleles were not reported by Pemberton *et al.*, 2008). As the data of Pemberton *et al.* (2008) are based on a set of individuals that overlaps with that of Jakobsson *et al.* (2008), this experiment mimics the scenario in which a genotyping chip is used on a set of samples and imputation of additional genotypes at marker positions that were not previously typed in the same samples is of interest (Figure 2.1d). This scenario occurs, for instance, in meta-analyses of multiple GWA studies (Barrett *et al.*, 2008; Lettre *et al.*, 2008; Loos *et al.*, 2008; Zeggini *et al.*, 2008).

In addition to reporting the proportion of alleles estimated correctly as the measure of imputation accuracy, we also calculated the square of a linear correlation coefficient between the imputed and directly-measured genotypes. At each SNP for which the true genotypes were masked, we coded the possible genotypes as 0, 1, or 2, representing the possible counts of the minor allele at this SNP in the target population. Let x_i denote the imputed genotype for individual i, and let $\bar{x}$ denote the mean value of the imputed genotypes across individuals. Similarly, let g_i and $\bar{g}$ denote the analogous quantities for the true genotypes. Then, the statistic, r^2, is computed as

$$r^2 = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})(g_i - \bar{g})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (g_i - \bar{g})^2}} \right)^2,$$

where n is the number of individuals in the population sample. This squared correlation coefficient was then averaged across SNPs to obtain a summary measurement for each population.

2.3 Results

2.3.1 Inferring Missing Genotypes without Additional Reference Individuals

Imputation accuracies for each population, as a function of the proportion of missing data, are displayed in Figure 2.2. Here, no reference panel has been used, and we observe a decrease in accuracy with an increasing proportion of missing data. The Pima and Colombian groups exhibited the highest imputation accuracies (>92% with 15% of genotypes missing). Across populations, the degree to which the proportion of missing genotypes affects imputation accuracy is relatively constant, as is evident in the parallel trajectories across populations in the figure. Over the range of missing data proportions examined, we did not observe a qualitative difference in population rankings by imputation accuracy. Populations from the Americas and Oceania had the highest imputation accuracy, followed by populations from Asia and Europe; African populations had the lowest imputation accuracy. Because the choice of the proportion of missing genotypes had relatively little influence on population rankings by imputation accuracy, especially for proportions less than ∼30%, we proceeded to subsequent analyses with a single proportion of missing genotypes equal to 15%.

Figure 2.3 shows the relationship between imputation accuracy and sample size, when unknown genotypes were imputed based only on information from within a population sample (i.e., without a reference panel). The imputation accuracy, as measured by the proportion of alleles imputed correctly, increases as sample size increases. The pattern across populations is similar to that in Figure 2.2, with popu-

lations from the Americas and Oceania having the highest imputation accuracy and African populations having the lowest imputation accuracy. The boost in accuracy provided by increasing the sample size is greatest when the sample size is small.

To assess the importance of the particular sequence of individuals employed in evaluating the role of sample size, for each population sample we used an additional random ordering of individuals. Figure S2.1 shows the imputation accuracy as a function of sample size in the absence of a reference panel for each of two sets of permuted samples. The point-wise differences between the values in the two plots in Figure S2.1 are shown in Figure S2.2, which displays no systematic difference in imputation accuracy as a function of sample size between the two permuted samples. The maximal difference in imputation accuracy between the two permuted samples was less than 0.5% in most populations. Consequently, the impact of using a particular sequence of individuals in the evaluation of imputation accuracy appears to be minimal.

2.3.2 Inferring Missing Genotypes with Additional Reference Individuals

2.3.2.1 Imputation Accuracy versus Panel Size

Figure 2.4 shows the relationship between imputation accuracy, based on each of the three HapMap reference panels, and the size of the panels. In the first three columns, we plot the imputation accuracy from inference of missing genotypes in each population, based on a single HapMap panel. In the final (rightmost) column, we plot the maximal imputation accuracy for each population, taken point-wise from the first three columns. Generally, when we used a single HapMap reference panel, higher imputation accuracies occurred in populations from the same geographic region as the reference panel and lower imputation accuracies occurred in African populations. With the YRI sample as the reference panel, both the highest and lowest imputation accuracies occurred in populations from Africa (Yoruba and San, respectively).

We generally observed increasing imputation accuracy with increasing reference panel size. Averaging across all 29 populations and all three HapMap reference panels, the increase in imputation accuracy was 3.21% when the reference panel size increased from 10 to 20 haplotypes; for subsequent additions of 10 reference haplotypes, the associated increases were 1.06%, 0.56%, 0.35%, 0.23%, 0.18%, 0.13%, 0.11%, 0.10%, 0.07%, and 0.06%. When we used the HapMap CEU or CHB+JPT sample as the reference panel, the imputation accuracy appeared to reach a plateau as the reference panel size approached 120 haplotypes. However, we did not observe as clear a plateau when using the HapMap YRI sample as the reference panel, particularly for the Yoruba HGDP sample.

When we considered the maximal imputation accuracy attained by using a single HapMap reference panel of size 120 haplotypes, European populations generally had the highest accuracy, followed by populations from East Asia, Central/South Asia, the Americas, the Middle East, Oceania, and Africa (Figure 2.4). The maximal imputation accuracies of populations within a geographic region displayed more variation in Africa and the Middle East than in other geographic regions. For example, when using 120 haplotypes from the reference panel, we found that African and Middle Eastern populations had a wider range of maximal imputation accuracies (9.8% for African populations and 2.8% for Middle Eastern populations) than, for instance, the Central/South Asian populations (<1% between the highest and lowest accuracies).

Figure 2.5 summarizes with a bar plot the maximal imputation accuracy achieved by one of the HapMap reference panels, each of size 120 haplotypes, for each population. The colors of the bars indicate which HapMap panel was utilized to produce the maximal imputation accuracy. In African populations, we obtained the maximal imputation accuracy using the HapMap YRI sample as the reference panel. Populations from Europe, Central/South Asia, and the Middle East as well as Maya from the Americas attained their maximal imputation accuracies using the HapMap

CEU panel, whereas populations from East Asia and Oceania as well as Pima and Colombian from the Americas achieved their maximal accuracies with the HapMap CHB+JPT reference panel.

2.3.2.2 Imputation Accuracy versus Panel Composition

For each population, Figure 2.6 displays the imputation accuracy based on mixtures of HapMap reference panels, indicating with a darkened circle the mixture of HapMap samples that produced the maximal imputation accuracy. The vertices of a triangle in Figure 2.6 represent imputation accuracies based solely on a single HapMap group, and the interior points represent imputation accuracies achieved by using mixtures of HapMap reference haplotypes (see Materials and Methods). The colors correspond to the nine quantiles of the observed imputation accuracies across all mixtures and all populations, with darker colors representing higher imputation accuracies. Each point in a triangle is colored according to the imputation accuracy produced by the panel mixture corresponding to the point.

With only a few exceptions, the panel mixture that led to the maximal imputation accuracy for a particular population had as its primary component the same HapMap reference panel that produced the maximal imputation accuracy individually in Figure 2.5. Specifically, the YRI panel was the primary component of the mixture for all African populations, the CEU panel was the primary component for all European populations, and the CHB+JPT panel was the primary component for populations from East Asia, Oceania and the Americas. However, populations from the Middle East and Central/South Asia did not display such homogeneous patterns for the major contributing HapMap panel in the optimal mixture. In two Middle Eastern groups, Mozabite and Bedouin, the HapMap YRI and CEU samples contributed equally to their optimal mixtures of reference haplotypes, while in the other two Middle Eastern groups, Palestinian and Druze, the CEU sample alone served as the

major contributing HapMap reference panel. For populations from Central/South Asia, the major contributing HapMap panels were the CEU sample in the Balochi group and the CHB+JPT sample in the Kalash and Uygur groups; the optimal mixture for the Burusho group contained equal contributions from the HapMap CEU and CHB+JPT samples.

Compared with imputation accuracy obtained using only a single HapMap reference group (Figures 2.4 and 2.5), in 23 of 29 populations, the major contributing HapMap sample in the mixtures that produced the maximal imputation accuracies corresponded to the single highest-accuracy panel in the analysis of HapMap panels individually. In the Kalash, Uygur, and Maya populations, the major contributing HapMap samples differed from the samples that produced the highest imputation accuracy when we evaluated HapMap panels separately; the Mozabite, Bedouin, and Burusho populations each had two HapMap panels contributing the same number of reference haplotypes in the optimal mixtures.

When we considered imputation accuracy across populations based on the 36 mixtures of reference panels, European and East Asian populations had generally higher imputation accuracies that fell within the top quantiles. With the exception of the Yoruba population, African populations had substantially lower imputation accuracies that mostly fell within the bottom quantiles. The highest imputation accuracy across all points in Figure 2.6 was 97.83% in the Basque population (based on a mixture consisting of 48 CHB+JPT haplotypes, all 120 CEU haplotypes, and no YRI haplotypes). The lowest imputation accuracy accuracy among all points tested—the minimum value across all 29×36 choices of a population sample and a reference panel—was 78.20%, in the San population (based on the entire CHB+JPT panel, 180 haplotypes). While the use of mixed reference panels resulted in increased imputation accuracy in all populations, the choice of all 210 HapMap individuals as the reference panel did not yield the highest imputation accuracy in any of the 29

populations. However, this choice generally produced similar imputation accuracy to that of the optimal mixture; across populations, the mean difference between imputation accuracy based on the optimal mixture and that based on the full HapMap sample was 0.0059. This value was less than the mean difference between imputation accuracy based on the optimal mixture and that based on the optimal vertex (0.0079).

2.3.3 Application to Untyped Markers

Figure 2.7 and Table S2.1 present imputation accuracy for inference of unknown genotypes in the untyped chromosome 21 markers of Jakobsson *et al.* (2008), based on individual HapMap panels and on mixtures of two or three HapMap panels. As indicated by the bar plot in Figure 2.7, five of seven combinations of HapMap panels produced the highest imputation accuracy in at least one population. The two combinations that did not serve as the optimal reference panel in any of the populations were the HapMap CEU sample and the combination of the YRI and CHB+JPT samples. Except in five groups (San, Mbuti Pigmy, Yoruba, Mandenka, and Lahu), most populations we examined benefited from use of a combination of two or more HapMap samples as the reference panel to impute genotypes at untyped markers on chromosome 21. The highest maximal imputation accuracy was 96.05%, occurring in a European population, Adygei, and the lowest maximal imputation accuracy was 89.12%, occuring in an African population, San.

In this setting, where mixtures of HapMap panels are coarser than those displayed in Figure 2.6, for 11 of 29 populations the imputation accuracy was the highest when we constructed the reference panel from all available HapMap individuals. Seven of these 11 groups represent populations of Eurasia with some degree of dissimilarity from the HapMap groups in northern and western Europe and in China and Japan; the other four are from Oceania and the Americas.

We obtained comparable results for the choice of reference panel when, in place of

imputation accuracy, we considered the squared correlation of imputed and measured genotypes, r^2, as a measure of the performance of the genotype imputation procedure (Figure 2.8 and Table S2.2). Unlike in Figure 2.7, however, populations from the Americas had the highest values of r^2. Across populations, the highest maximal r^2, 0.9618, occurred in the Pima population, and the lowest maximal r^2, 0.7397, occurred in the Mbuti Pygmy population. Among the seven combinations of the HapMap panels, the CHB+JPT sample was the only panel that did not serve as the optimal panel for any of the populations. In 25 of 29 populations, the use of two or three HapMap samples produced the maximal r^2 between the imputed genotypes and those directly measured by Conrad *et al.* (2006). A single HapMap panel (YRI) produced the highest r^2 in San, Yoruba, and Mandenka; another individual panel (CEU) produced the highest r^2 in the Russian population. When we used all available HapMap individuals as the reference panel, we obtained the maximal r^2 in 9 populations, 8 of which were among the 11 populations for which imputation accuracies were the highest when using the full HapMap set in Figure 2.7.

2.4 Discussion

Until now, nearly all imputation-based GWA studies have been performed in populations of European descent. As genotyping costs decrease, such studies will likely begin to include individuals from an increasing diversity of populations. Due to the success of recent studies that have leveraged external reference samples for imputation of unmeasured genotypes and due to the potential we have demonstrated for accurate genotype imputation in diverse populations, it is likely that the imputation approach can be applied successfully to GWA studies in which the sampled individuals are more distantly related to the samples that comprise available reference panels. This investigation can therefore serve as an initial resource for the design and analysis of imputation-based GWA studies in these diverse populations.

We characterized the levels of LD in 29 HGDP populations using the practical metric of imputation accuracy, the ability to estimate missing genotypes based on patterns of LD. Although our evaluations of imputation accuracy based on the HGDP samples alone (without using a reference database) are somewhat constrained by the small sample sizes, we obtained relative imputation accuracies among the HGDP populations that reflect previously observed levels of LD. For example, these imputation accuracy comparisons correspond closely to the pairwise LD calculations described by Jakobsson *et al.* (2008). Indeed, the Spearman correlation coefficient of population rankings by imputation accuracy at 15% missing data (Figure 2.2) and population rankings by the pairwise LD statistic r^2 for markers at 10 kb distance (Figure S4 of Jakobsson *et al.*, 2008) was 0.9680 (Tables 2.1 and 2.2).

Our assessments of which reference panels are most appropriate for imputation in different populations are reminiscent of evaluations of tag SNP portability in the same populations (Conrad *et al.*, 2006; González-Neira *et al.*, 2006; Pemberton *et al.*, 2008). When considering the three HapMap samples separately, in nearly all populations, we obtained the maximal imputation accuracies for the data of Conrad *et al.* (2006) and Pemberton *et al.* (2008) using the same HapMap groups that produced the highest proportion of variation tagged (PVT) as reported by these studies. The only exception was the Mozabite population, in which the CEU panel achieved the highest imputation accuracy and the YRI panel achieved the highest PVT. These results nonetheless were compatible as both optimal mixtures of HapMap samples in Mozabites—the one that produced the highest imputation accuracy and the one that produced the highest PVT (Pemberton *et al.*, 2008)—contained equal proportions of the HapMap CEU and YRI panels.

More generally, we observed a notable consistency in the PVT and imputation accuracy results for mixture reference panels. In 24 of 29 populations, the major contributing HapMap group in the optimal mixture for the purpose of genotype im-

putation (Figure 2.6) corresponded to the major group in the optimal mixture for tag SNP selection (Pemberton *et al.*, 2008). In the Burusho population from Central/South Asia, the optimal mixture for imputation contained equal HapMap CEU and CHB+JPT components, whereas the CEU panel alone served as the major contributing HapMap group in the optimal mixture for tag SNP selection (Pemberton *et al.*, 2008). In the other four populations (Uygur and Kalash from Central/South Asia and Colombian and Maya from the Americas), the major contributing HapMap group was the HapMap CHB+JPT panel in the optimal mixture for imputation and the CEU panel in the optimal mixture for selecting tag SNPs.

Caution needs to be exercised in comparing imputation accuracy results from our study with tag SNP results from Conrad *et al.* (2006) and Pemberton *et al.* (2008). In our evaluation of the effect of panel size on imputation accuracy using individual HapMap panels (Figure 2.3), we adjusted for differences in panel size by studying HapMap samples of equal size (120 haplotypes), whereas in assessing the potential of mixture panels to infer unknown genotypes (Figure 2.6), we utilized up to 180 haplotypes from the CHB+JPT reference group to allow for the use of all available HapMap samples. Pemberton *et al.* (2008), on the other hand, used subsets of the CHB+JPT panel of size 120 haplotypes throughout their mixture analyses. Our decision to utilize the HapMap CHB+JPT panel in its entirety could in part explain the increased utility of the CHB+JPT panel in the optimal mixtures for the five aforementioned Central/South Asian and American populations.

Although LD levels predicted imputation accuracy extremely well when we imputed genotypes without reference panels, with reference panels, LD levels were less predictive of imputation accuracy (e.g., Tables 2.1 and 2.2, Spearman correlation coefficient of 0.5795 between the maximal imputation accuracy in Figure 2.6 with the pairwise LD statistic, r^2, at 10 kb). African populations, whose levels of LD were generally quite similar (Jakobsson *et al.*, 2008), varied considerably in imputation ac-

curacy, with the highest values occurring in the lower-LD Yoruba population and the lowest values occurring in the higher-LD Mbuti Pygmy and San populations. Instead of being highest for populations from the Americas and Oceania, who exhibit the highest LD levels, in most analyses imputation accuracy was highest for European and East Asian populations closely related to populations from the reference panels. When the squared correlation coefficient between imputed and measured genotypes was used as the measure of imputation performance, however, the rankings of populations matched the pattern expected based on LD levels somewhat more closely (Tables 2.1 and 2.2).

The accuracy with which genotypes can be imputed using a reference panel is a function of multiple factors, including the similarity of haplotypes in the study sample and reference panel, as well as the allele frequencies and levels of LD in the study sample. For most populations in which imputation accuracy was high, the high value might have been expected on the basis of at least one of these factors. For the Basque population, who had the highest imputation accuracy in some analyses, a lower imputation accuracy might have been expected due to the status of the population as a linguistic isolate. However, previous analyses of the same samples have found this population to be genetically similar to other European populations, with similar levels of LD (Jakobsson *et al.*, 2008; Li *et al.*, 2008), so that a similar imputation accuracy for Basques and other European populations is not surprising. Another factor that could have contributed to high imputation accuracy in Basques and other Europeans is the possibility that European reference haplotypes might have been estimated more accurately than East Asian reference haplotypes, due to the availability of offspring in trios. Finally, the properties of the markers studied in the HapMap reference samples might influence imputation accuracy; many of the markers used were likely chosen to be informative about LD in Europeans, potentially leading to increased imputation accuracy in European populations.

Here we have not extensively examined the ability of LD-based algorithms to impute genotypes at SNPs of specific allele frequencies. Our data do, however, permit a preliminary investigation of the effect of allele frequency on imputation accuracy in different populations. For each population, Figure 2.9 compares imputation accuracy for untyped markers with minor allele frequency (MAF) greater than 0.2 and untyped markers with MAF$\leq$0.2. In all 29 populations, the genotypes of markers in the lower-MAF category were imputed with fewer errors. African populations showed a high variability in the difference in imputation accuracy between lower-MAF and higher-MAF markers (Figure S2.3), with a difference as high as 8.2% in the San population. In most non-African populations, genotypes of higher-MAF markers were imputed almost as accurately as were those of lower-MAF markers—most notably in the Mozabite population, for whom the difference in imputation accuracies was only 0.3%. These observations are due, in part, to the distributions of allele frequencies at the imputed SNPs; populations whose MAF$>$ 0.2 and MAF$\leq$ 0.2 markers had a larger difference in mean minor allele frequency (Table S2.3) tended to display larger differences in imputation accuracy between the two SNP sets. A larger reference panel size will be of some help in increasing the potential for accurate imputation; the extent to which rare alleles are satisfactorily imputed will be more easily tested in projects that include larger reference sample sizes and, consequently, that include rarer alleles.

An examination of reference panel size could assist in characterizing the way in which imputation accuracy changes for alleles in different frequency categories as reference panels are enlarged; we note however that our analysis of imputation accuracy and reference panel size is restricted to the marker sets directly measured in the genome scan itself, whereas in practice, the accuracies of all imputed SNPs would be of interest. Because they were included on a commercial SNP chip, the SNPs available for testing are tag SNPs that have a somewhat regular spacing. If alleles

at a tag SNP are masked, then the distance to the nearest tag SNPs from which the imputation is performed might be greater than the corresponding distance for a randomly chosen SNP. Additionally, tag SNPs tend to have higher allele frequencies, at least for the populations in which the SNPs were discovered and the populations for which the chips were designed. Conclusions about the value of larger reference panels should be interpreted in this light and might potentially benefit from results obtained in simulations (Pei *et al.*, 2008).

In evaluating genome-wide imputation accuracies, results from rare SNPs are hidden by the large number of testable genotypes at SNPs with more frequent minor alleles. Furthermore, assessment of imputation accuracy of heterozygous genotypes at rare SNPs is obscured by the imputation accuracy statistic we use here. For instance, a procedure that always imputes the major allele will, on average, achieve 99.9% accuracy at a SNP with minor allele frequency of 1/1000. However, this high level of accuracy can hide a high error rate for individuals with the rare allele. As detection of rare alleles and their interactions becomes more feasible in association studies, it will be of interest to more carefully assess the accuracy with which rare alleles can be imputed.

We note that whereas our investigations that did not rely on a reference panel were affected by the sizes of the HGDP samples, our evaluations of imputation accuracy when utilizing reference panels were not strongly dependent on sample size. This result is due to the manner in which we conducted our investigations, which was motivated by current strategies for imputation-based mapping in GWA studies. Specifically, conditional on the reference haplotypes, we analyzed the study samples independently rather than including other study individuals when imputing genotypes of each particular study individual. Therefore, average imputation accuracies reported here are unbiased estimates of what would be obtained from studying the entire population, provided that the individuals chosen were sampled randomly from

the population.

Because of the conditional independence of study individuals during the analysis (given the reference haplotypes), the scheme we used to evaluate optimal mixtures (e.g. Figure 2.6) also mimicked the current setting for analyses of GWA data, where the information for imputing a single unobserved genotype comes entirely from the reference panel. Although for this particular investigation, we did not force all genotypes to be unobserved at specified loci and instead masked individual genotypes completely at random, our imputation accuracy results using genotypes masked at random in Figures 2.4-2.6 were similar to those using completely untyped markers in Figures 2.7 and 2.8. Results from our detailed investigation of optimal mixtures might therefore serve as a basis for methods that appropriately weigh reference samples from the various panels while utilizing all available information.

An alternative approach to evaluating optimal reference panel composition, which we did not pursue, is to identify the mixture that produced the maximal imputation accuracy among mixtures of a fixed panel size, in order to more thoroughly evaluate the maximal imputation accuracy as a function of reference panel size. This approach is constrained by the difference in the HapMap reference sample sizes, so it cannot consider a mixture sample larger than 120 haplotypes (60 individuals), the smallest HapMap reference panel size. Thus, taking into consideration the effect of reference panel size on imputation accuracy (Figure 2.4), our use of the largest mixed sample permitted by a given ratio is motivated by the goal of imputing based on as many reference individuals as possible, given currently available databases. Although the optimal mixtures in Figure 2.6 for the 29 populations were not comprised of all 420 haplotypes (from 210 unrelated HapMap individuals), in many cases the difference between the maximal accuracy and that obtained from using all haplotypes was relatively small and, for such populations, the collection of all haplotypes would form a convenient reference.

2.5 Web Resources

HapMap phase II data,

`http://ftp.hapmap.org/phasing/2006-07_phaseII/phased/`

MACH software, `http://www.sph.umich.edu/csg/abecasis/mach/`

Seattle SNPs Variation Discovery Resource, `http://pga.gs.washington.edu`

2.6 Appendix

2.6.1 Software Settings

The options implemented in MACH that we used included mle, mldetails, interimInterval, rounds, errorRate, compact, greedy, autoFlip, and mask. The first two options generate SNP-specific information (e.g., marker name, allele labels, minor allele frequency, etc.) as well as genotype-level maximum likelihood estimates of genotypes, allele dosage, confidence scores, and posterior probabilities for the three possible genotypes; interimInterval outputs intermediate imputation results; rounds specifies the number of runs for the Markov sampler (set to 20); errorRate provides to the algorithm an omnibus measure reflecting a combination of genotyping error, gene conversion, recurrent mutation, and assay inconsistencies between multiple platforms or laboratories (set to 10^{-3}); compact reduces memory requirements at the cost of computational time; greedy treats the reference panel (and not the combination of the study and reference samples) as the only source of reference haplotypes; autoFlip switches the alleles at a given locus in the study samples to the complementary alleles when it is discovered that the reference panel uses alleles A and T and the study sample uses C and G (or vice versa). The mask option, used throughout our analyses except in the application to untyped markers, specifies the proportion of genotype data to be randomly masked for evaluation of imputation accuracy.

2.6.2 Obtaining Mixtures of HapMap Reference Panels

Here, we solve for the numbers of haplotypes, (a_1, a_2, a_3), that maximize the total number of haplotypes present when a ratio of integers $i_1 : i_2 : i_3$ is specified for the relative numbers of haplotypes in three groups.

Suppose that positive integers k and n are given, that i_j is an integer in $[0, k]$ for each j from 1 to n, and that $\sum_{j=1}^{n} i_j = k$. Suppose also that for each j from 1 to n, a positive integer A_j is given, and a_j is an integer in $[0, A_j]$. We aim to find $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ such that $\sum_{j=1}^{n} a_j$ is as large as possible and such that $a_1 : a_2 : \ldots : a_n = i_1 : i_2 : \ldots : i_n$.

Without loss of generality, suppose $i_1 \geq i_2 \geq \ldots \geq i_n$. Because $a_1 : a_2 : \ldots : a_n = i_1 : i_2 : \ldots : i_n$, $a_1 i_j / i_1$ must be an integer for each j. Because

$$\frac{a_1 i_j}{i_1} = \frac{a_1 i_j / \gcd(i_1, i_j)}{i_1 / \gcd(i_1, i_j)},$$

where gcd represents the greatest common divisor, for each j, a_1 must be a multiple of $i_1 / \gcd(i_1, i_j)$, as $i_j / \gcd(i_1, i_j)$ and $i_1 / \gcd(i_1, i_j)$ are relatively prime. It follows that a_1 is a multiple of $\mathrm{lcm}(i_1 / \gcd(i_1, i_2), \ldots i_1 / \gcd(i_1, i_n))$, where lcm represents the least common multiple. Considering that $a_j = a_1 i_j / i_1$ and $a_j \leq A_j$ for each j, $a_1 \leq \min(A_1, A_2 i_1 / i_2, \ldots A_n i_1 / i_n)$. As a result, the solution for a_1 in the vector $\mathbf{a}$ that maximizes $\sum_{j=1}^{n} a_j$ is

$$\begin{aligned} a_1 \;=\; & \mathrm{lcm}\left(\frac{i_1}{\gcd(i_1, i_2)}, \ldots, \frac{i_1}{\gcd(i_1, i_n)} \right) \times \\ & \left\lfloor \frac{\min(A_1, A_2 i_1 / i_2, \ldots A_n i_1 / i_n)}{\mathrm{lcm}(i_1 / \gcd(i_1, i_2), \ldots, i_1 / \gcd(i_1, i_n))} \right\rfloor. \end{aligned} \tag{2.1}$$

The other components of $\mathbf{a}$ are obtained using $a_j = a_1 i_j / i_1$.

In our analysis, $k = 7$, $n = 3$, and $(A_1, A_2, A_3) = (180, 120, 120)$. For each (i_1, i_2, i_3) with $i_1 + i_2 + i_3 = 7$ we obtain (a_1, a_2, a_3) using eq. 2.1. We chose $k =$

7 as this is the smallest value that permits use of the full HapMap, at the point $(i_1, i_2, i_3) = (3, 2, 2)$.

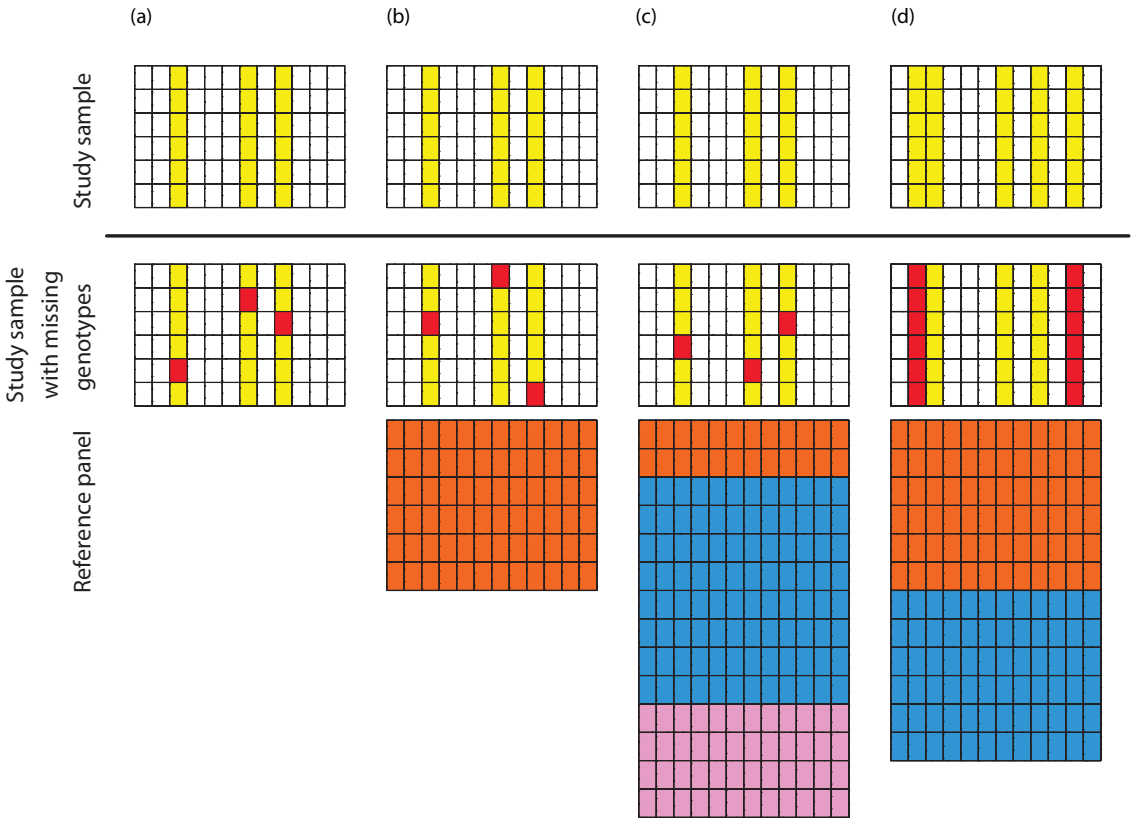

Figure 2.1: Schematic of experimental designs. The "Study sample" row represents data used to evaluate imputation accuracy in each design, with SNPs under consideration colored yellow. The "Study sample with missing genotypes" row represents corresponding data with the unknown genotypes that are imputed colored in red. The "Reference panel" row represents example reference panels based on which imputation of missing genotypes or genotypes of untyped markers is performed. In a data set, each row corresponds to a haplotype and each column corresponds to a SNP position. (a) Inference of missing genotypes without additional reference haplotypes. (b) Inference of missing genotypes with a reference panel of haplotypes from a single reference sample (either CEU, YRI, or CHB+JPT). (c) Inference of missing genotypes with a mixture reference panel, formed by taking a specified ratio of haplotypes from the HapMap CEU, YRI, and CHB+JPT samples. (d) Inference of genotypes of untyped markers with a mixture reference panel, formed by combining two or more HapMap samples. We evaluated imputation accuracy in (a)-(c) for randomly masked genotypes and in (d) for genotypes of untyped markers.

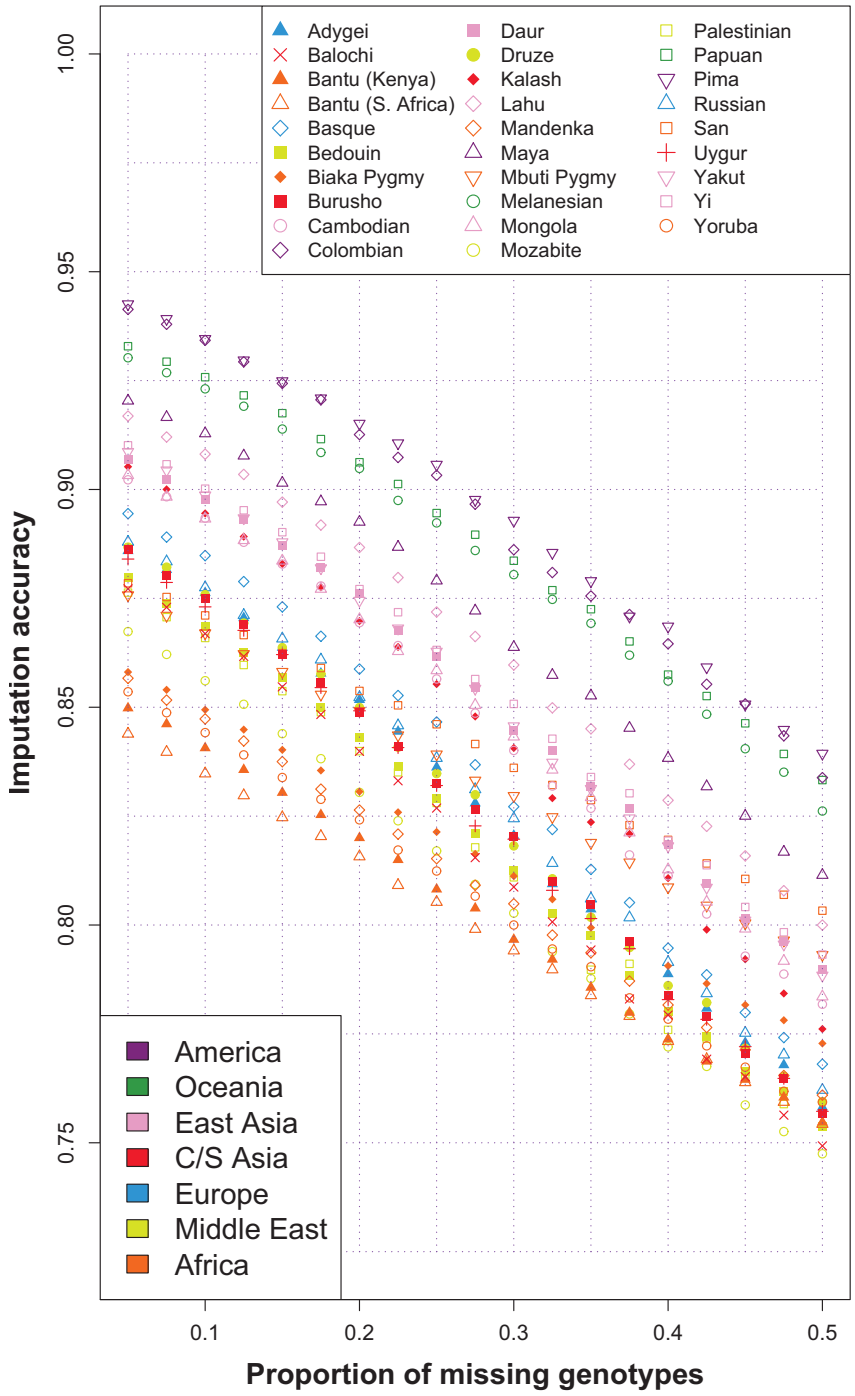

Figure 2.2: Imputation accuracy vs. proportion of missing genotypes, in each of 29 populations. This analysis was based on samples of six individuals per population, and did not use any reference panel.

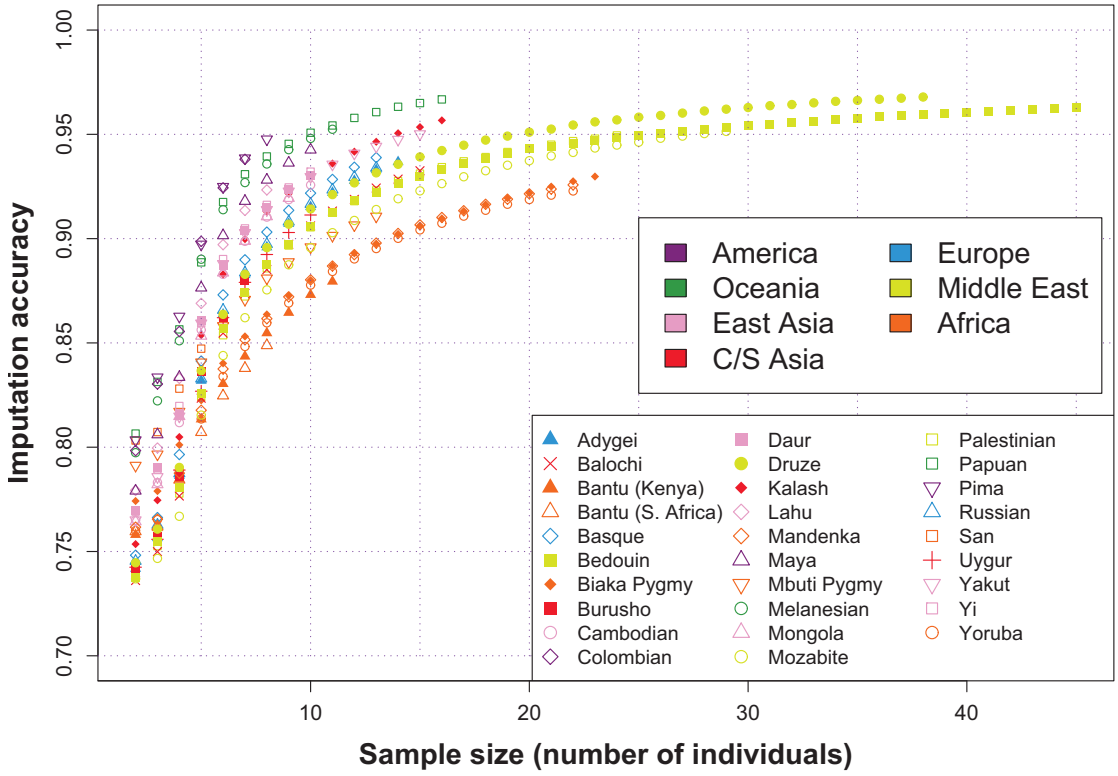

Figure 2.3: Imputation accuracy vs. sample size, in each of 29 populations. This analysis used a proportion of missing genotypes equal to 15%, and did not use any reference panel.

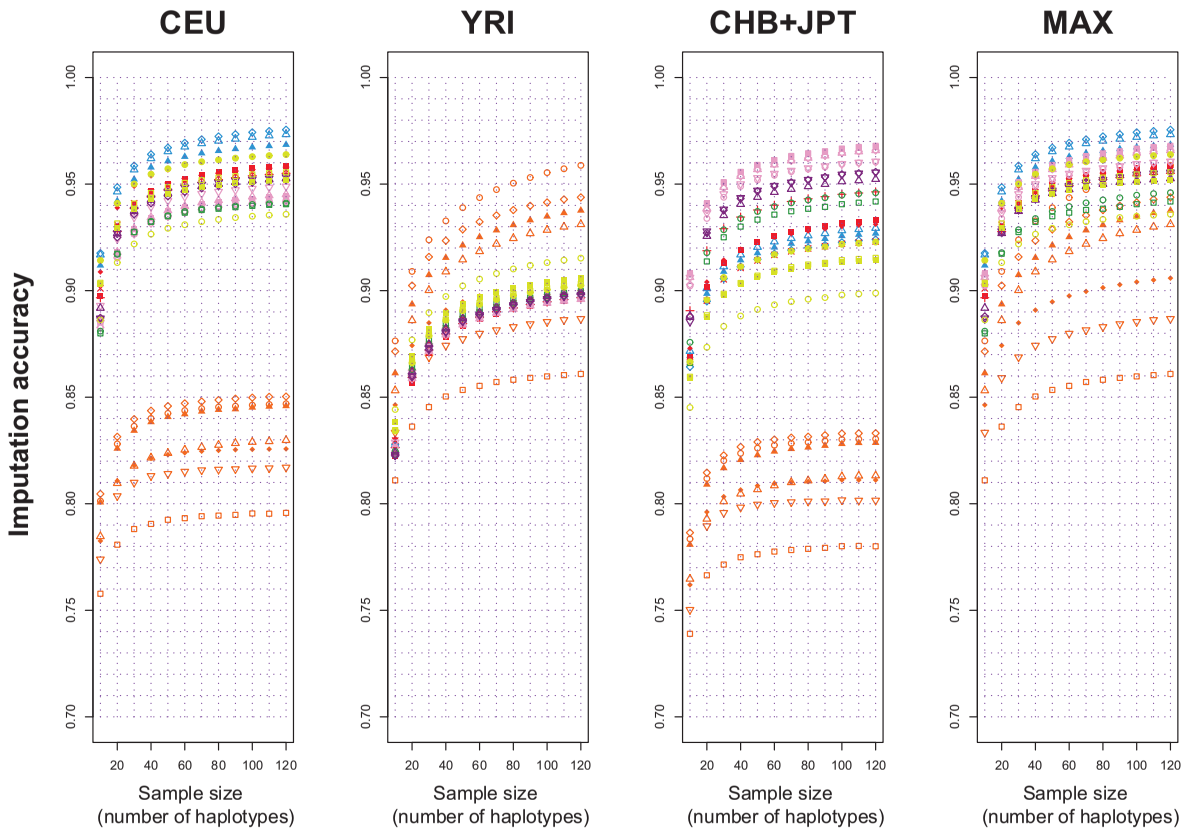

Figure 2.4: Imputation accuracy vs. reference panel size, in each of 29 populations, given a proportion of missing genotypes equal to 15%. To obtain comparable results, we used the entire HapMap YRI and CEU samples but only 120 of 180 HapMap CHB+JPT reference haplotypes. The rightmost column of "maximal" imputation accuracy represents the highest accuracy achieved by one of the HapMap reference panels, taken point-wise. Populations are color-coded and symbol-coded in the same manner as in Figure 2.3.

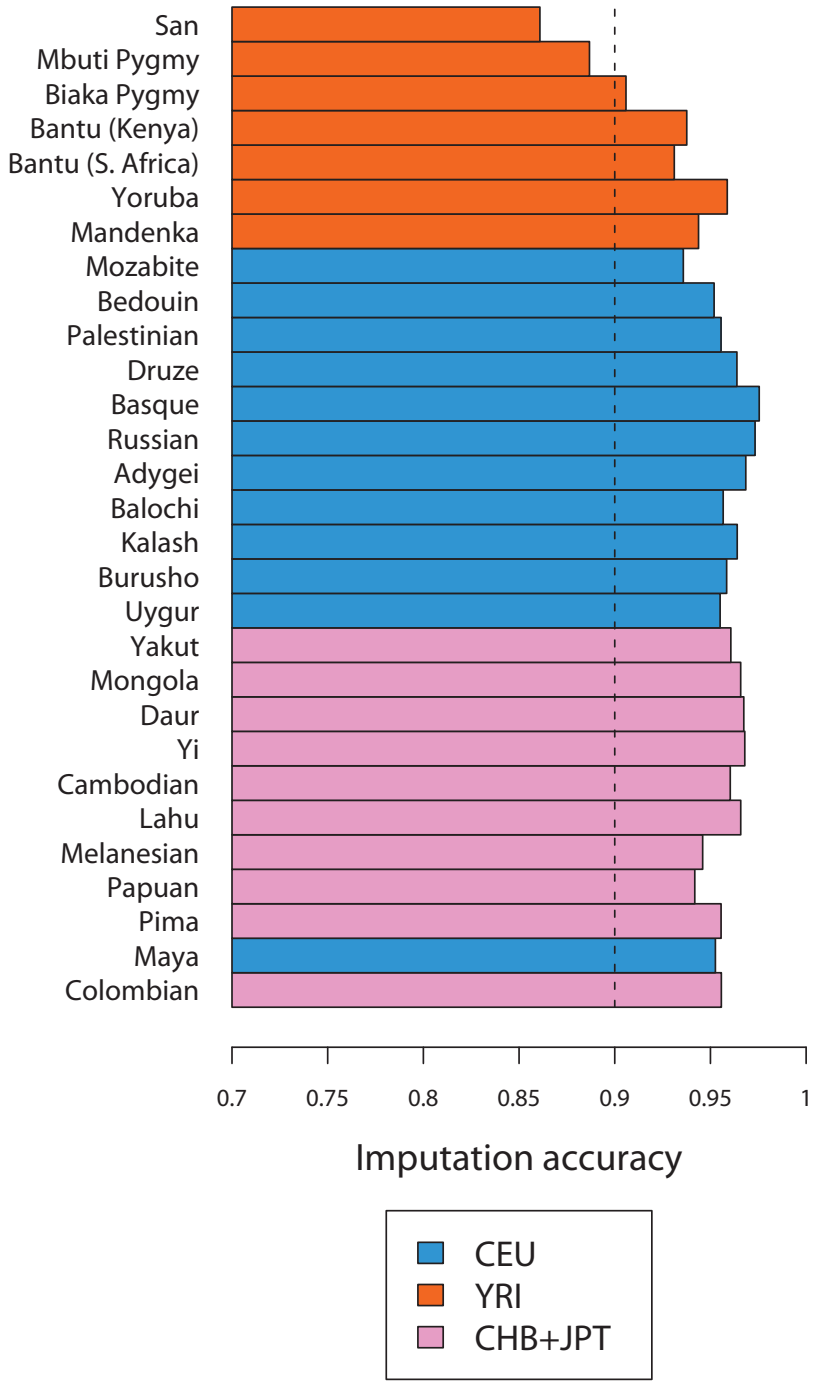

Figure 2.5: The maximal imputation accuracy achieved by one of the three HapMap reference panels, in each of 29 populations, given a proportion of missing genotypes equal to 15%. This plot corresponds to the imputation accuracy with a reference panel size of 120 haplotypes in the rightmost column (MAX) in Figure 2.4. For convenience in interpreting the figure, the vertical dashed line indicates 90% imputation accuracy.

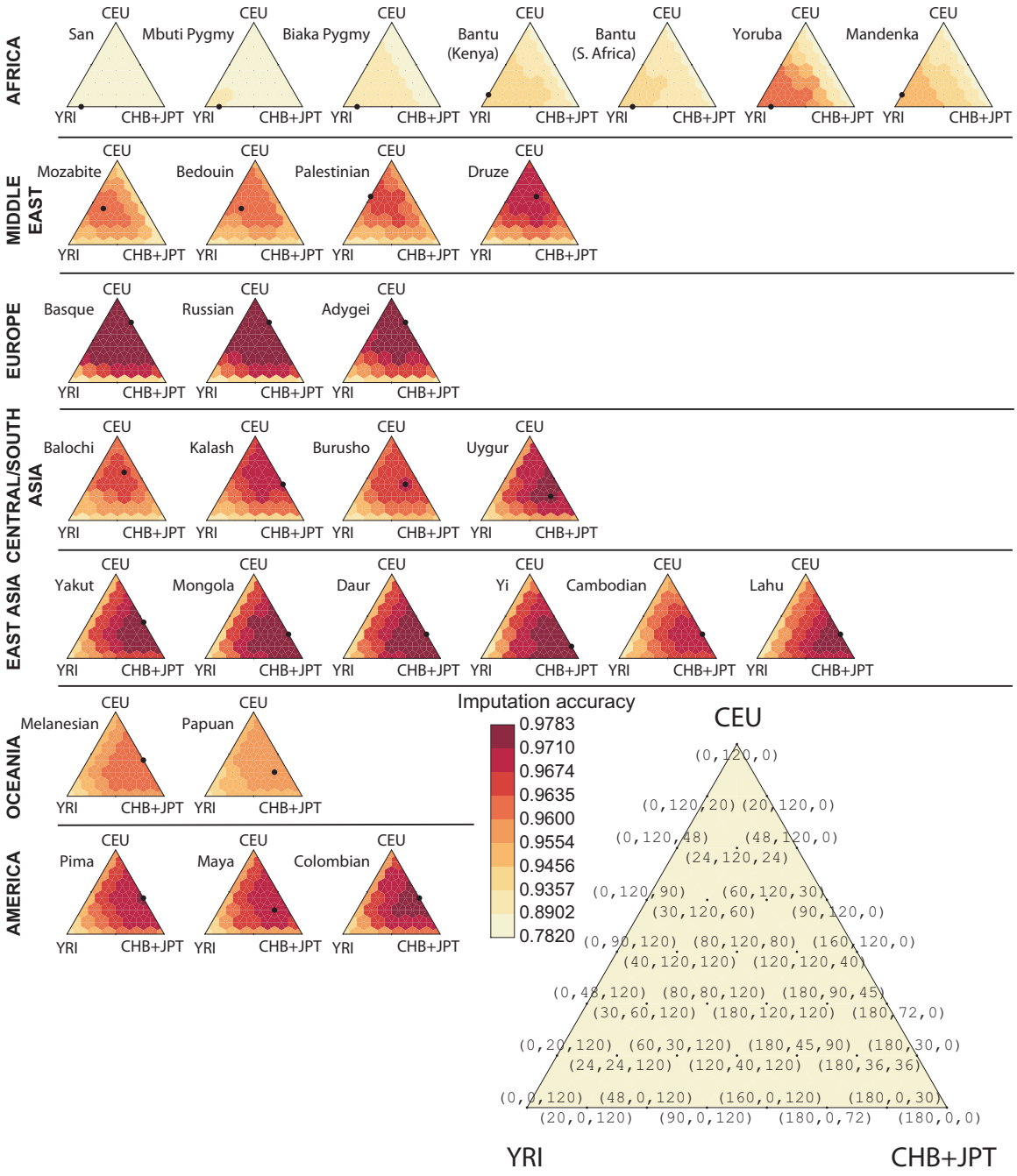

Figure 2.6: Imputation accuracy in each of 29 populations achieved by utilizing mixtures of HapMap samples chosen according to specified ratios. Each triangle presents imputation accuracy, for a given population, based on various mixtures of HapMap reference panels. The vertices of a triangle represent imputation accuracy based on a single HapMap group, while the edges and interior points represent imputation accuracy attained by using mixtures of HapMap reference panels. Darker colors indicate higher imputation accuracy; a darkened circle indicates the maximal imputation accuracy for a population.

Figure 2.6: (*figure caption continued from previous page*) The spacing of the cutoffs for the various colors was set so that across all 29 populations each color would be used equally often. The set of mixtures corresponded to the set of vectors (i_1, i_2, i_3) of nonnegative integers with $i_1 + i_2 + i_3 = 7$. For each vector, we used as the reference panel the largest possible mixture sample that consisted of a_1, a_2, and a_3 HapMap CHB+JPT, CEU, and YRI individuals, respectively, and that satisfied $a_1 : a_2 : a_3 = i_1 : i_2 : i_3$. Corresponding numbers of HapMap haplotypes in the mixtures, (a_1, a_2, a_3), are shown in the larger triangle. Imputation accuracy was evaluated using only chromosome 2, with a proportion of missing genotypes equal to 15%.

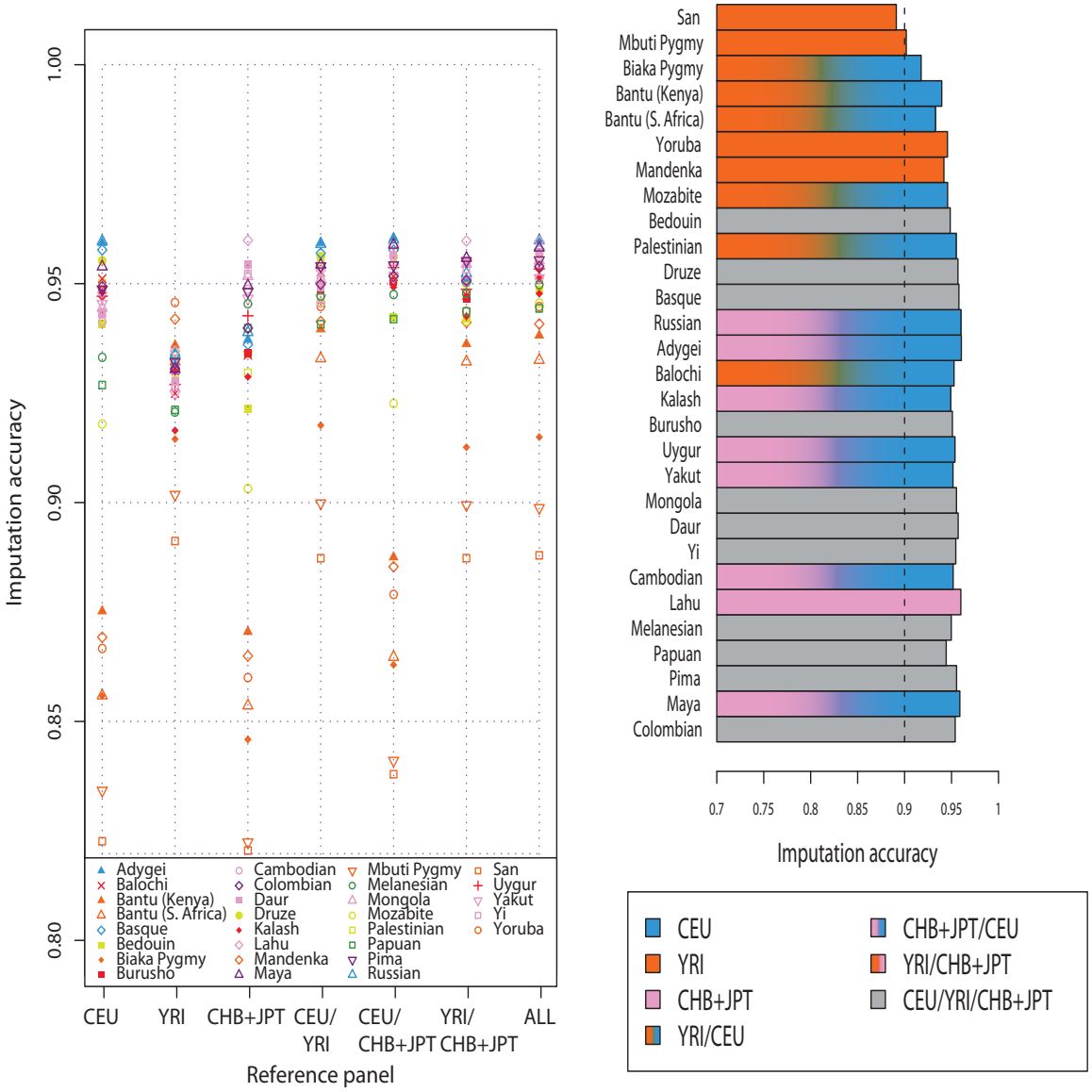

Figure 2.7: Imputation accuracy for inference of genotypes of untyped markers, based on any one or two or all three HapMap reference panels (with their original size). The plot on the left shows imputation accuracy based on each of seven choices. The bar plot on the right represents the maximal imputation accuracy among the seven choices, and is colored according to the choice of optimal reference panel. For convenience in interpreting the figure, the vertical dashed line indicates 90% imputation accuracy.

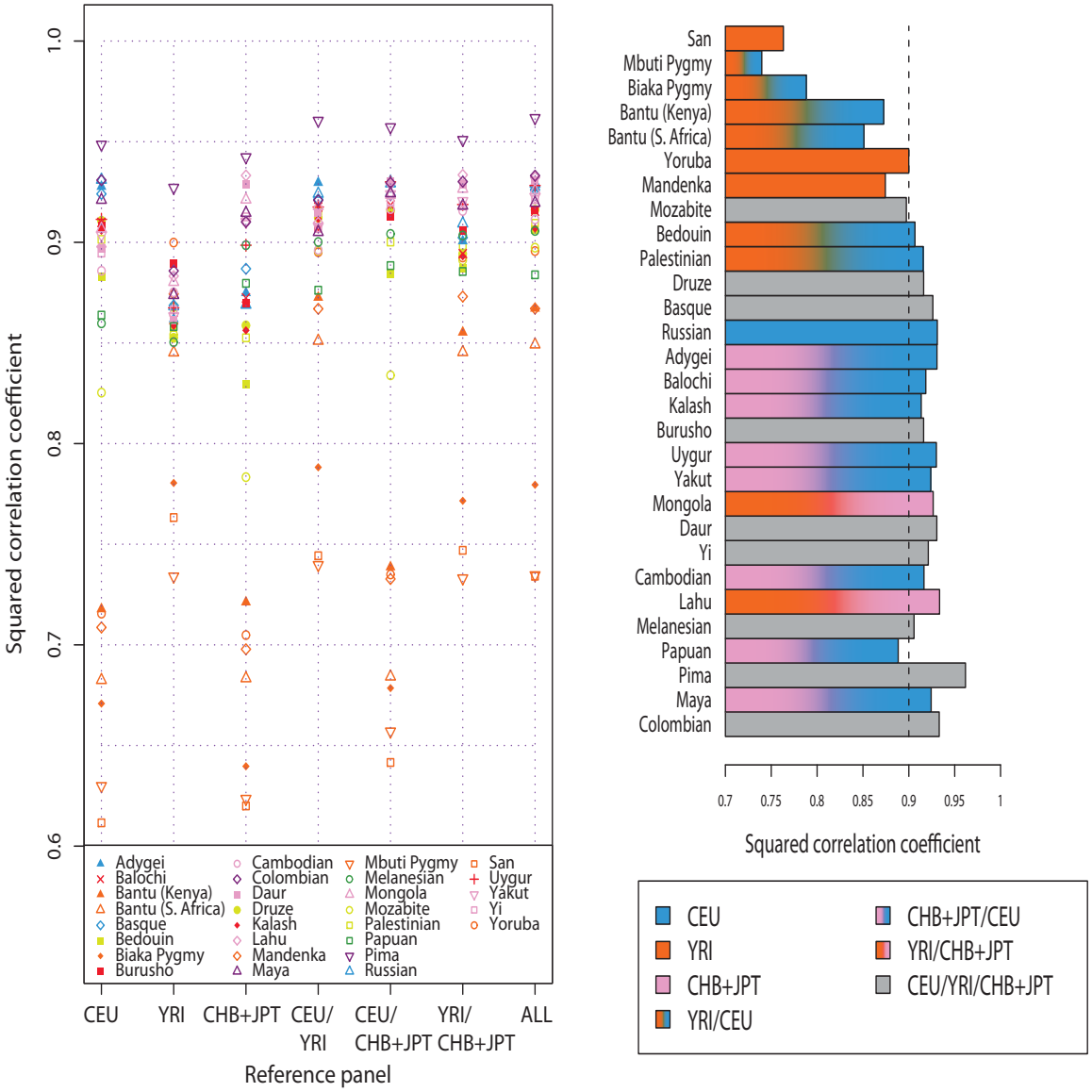

Figure 2.8: Squared correlation coefficient, r^2, between the genotypes imputed from the data of Jakobsson *et al.* (2008) and those directly measured in the data of Pemberton *et al.* (2008), based on any one or two or all three HapMap reference panels (with their original size). The plot on the left shows r^2 based on each of seven choices. The bar plot on the right represents the maximal r^2 among the seven choices, and is colored according to the choice of optimal reference panel. For convenience in interpreting the figure, the vertical dashed line indicates a squared correlation coefficient of 0.9.

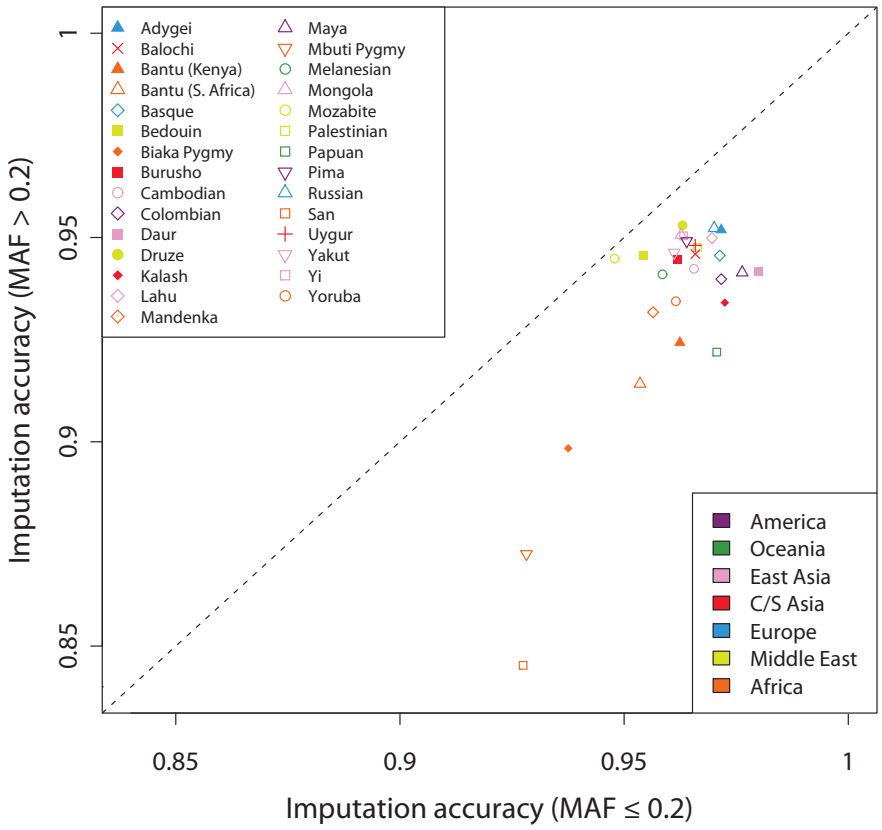

Figure 2.9: Imputation accuracy for genotypes at untyped markers in the data of Jakobsson *et al.* (2008) with minor allele frequency (MAF) greater than 0.2 vs. imputation accuracy for genotypes at untyped markers with MAF$\leq$0.2. For a given population, we separated markers into two categories based on their MAF in the population, on average placing 220 markers into the lower-MAF category and 293 into the higher-MAF category. Using the imputed genotypes described in Figures 2.7 and 2.8, for each of the seven reference panel choices, we determined the imputation accuracy, separately restricting our attention to low-MAF markers and to high-MAF markers. For each population, the highest of these seven numbers for the high-MAF markers is plotted on the y-axis and the highest of these seven numbers for the low-MAF markers is plotted on the x-axis (in some cases, the underlying optimal reference panel differed for the high-MAF and low-MAF markers). The diagonal dashed line indicates identical imputation accuracy for the two MAF categories. The difference between the imputation accuracy of the low-MAF markers and that of the high-MAF markers is plotted in Figure S2.3.

Table 2.1: Statistics compared across imputation scenarios

Scenario number	Figure displaying scenario results	Type of statistics	Description of imputation scenario
1	2	Imputation accuracy	15% randomly missing genotypes; imputation without reference panels
2	5	Imputation accuracy	15% randomly missing genotypes; imputation with the optimal single HapMap reference panel (among 3 choices)
3	6	Imputation accuracy	15% randomly missing genotypes; imputation with the optimal mixture HapMap reference panel (among 36 choices)
4	7	Imputation accuracy	Untyped markers; imputation with the optimal combination of HapMap reference panels (among 7 choices)
5	8	Squared correlation coefficient between imputed and measured genotypes	Untyped markers; imputation with the optimal combination of HapMap reference panels (among 7 choices)
6	S4 in Jakobsson *et al.* (2008)	Linkage disequilibrium statistic, r^2, at 10kb	N/A

Table 2.2: Spearman and Pearson correlation coefficients between measures of imputation accuracy in various scenarios

Scenario number	**1**	**2**	**3**	**4**	**5**	**6**
1		0.3910	0.5499	0.5217	0.6177	0.9680
2	0.3008		0.8852	0.8035	0.7453	0.4263
3	0.3755	0.9760		0.8744	0.8980	0.5795
4	0.3601	0.9699	0.9856		0.9034	0.5542
5	0.4405	0.9301	0.9653	0.9683		0.6507
6	0.9677	0.4225	0.5100	0.4971	0.5732	

For each scenario in Table 2.1, we obtained a list of values for the 29 populations, and the correlation coefficients between pairs among these lists are shown in the table. An entry in the table represents the correlation coefficient between lists for the scenarios in the appropriate row and column. The Spearman and Pearson correlation coefficients are shown in the upper and lower triangular areas on either side of the blank cells, respectively.

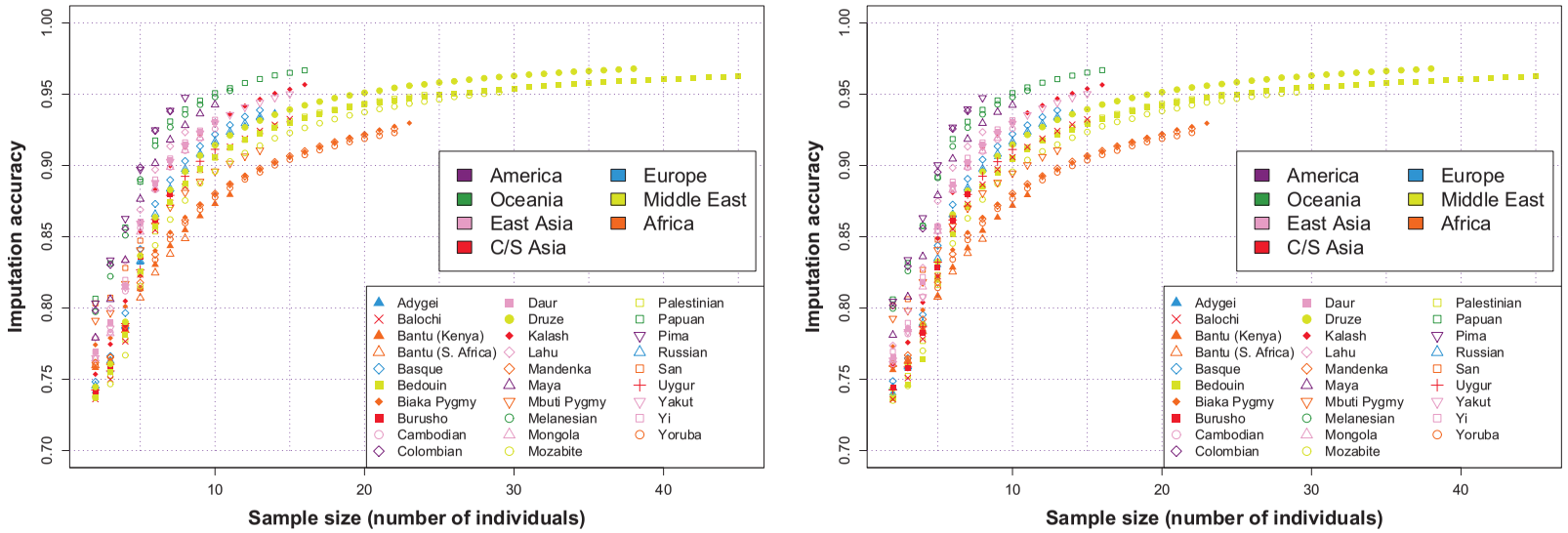

Figure S2.1: Imputation accuracy vs. sample size, in each of 29 populations. The two plots are based on two different subsets of individuals of the sample. The plot on the left is identical to Figure 2.3.

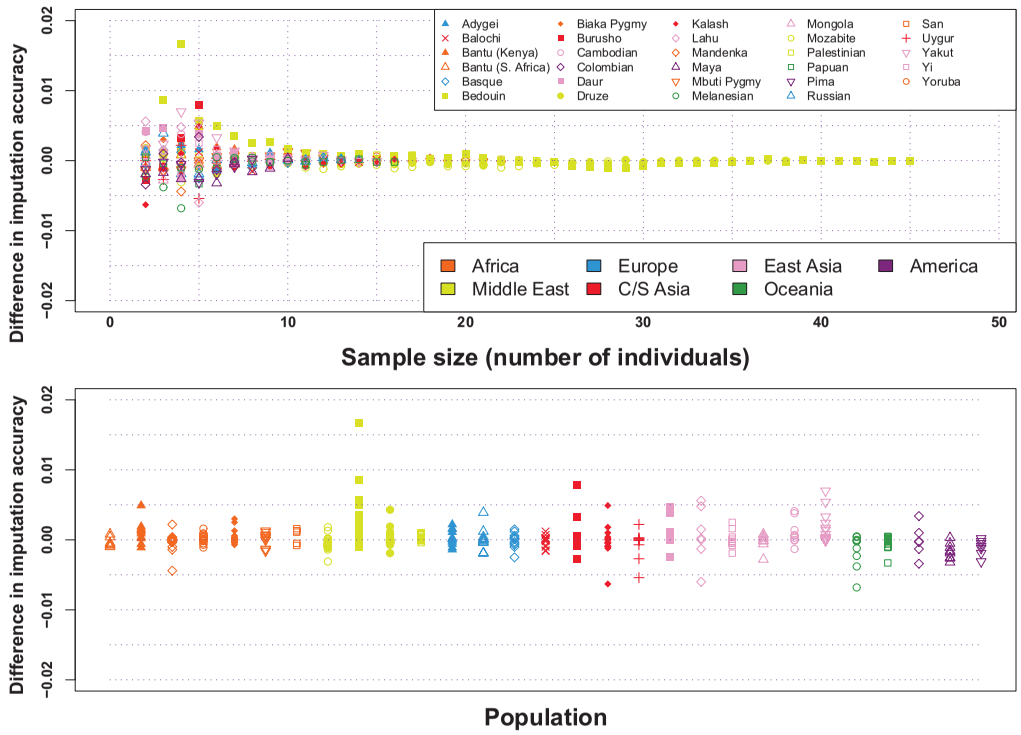

Figure S2.2: Difference in imputation accuracy assessed with one subset of individuals compared to a second subset based on another permutation of the individuals, in each of 29 populations. The points correspond to point-wise differences between the values in the two plots in Figure 2.7 (i.e., subtracting values in the right plot from corresponding values in the left plot).

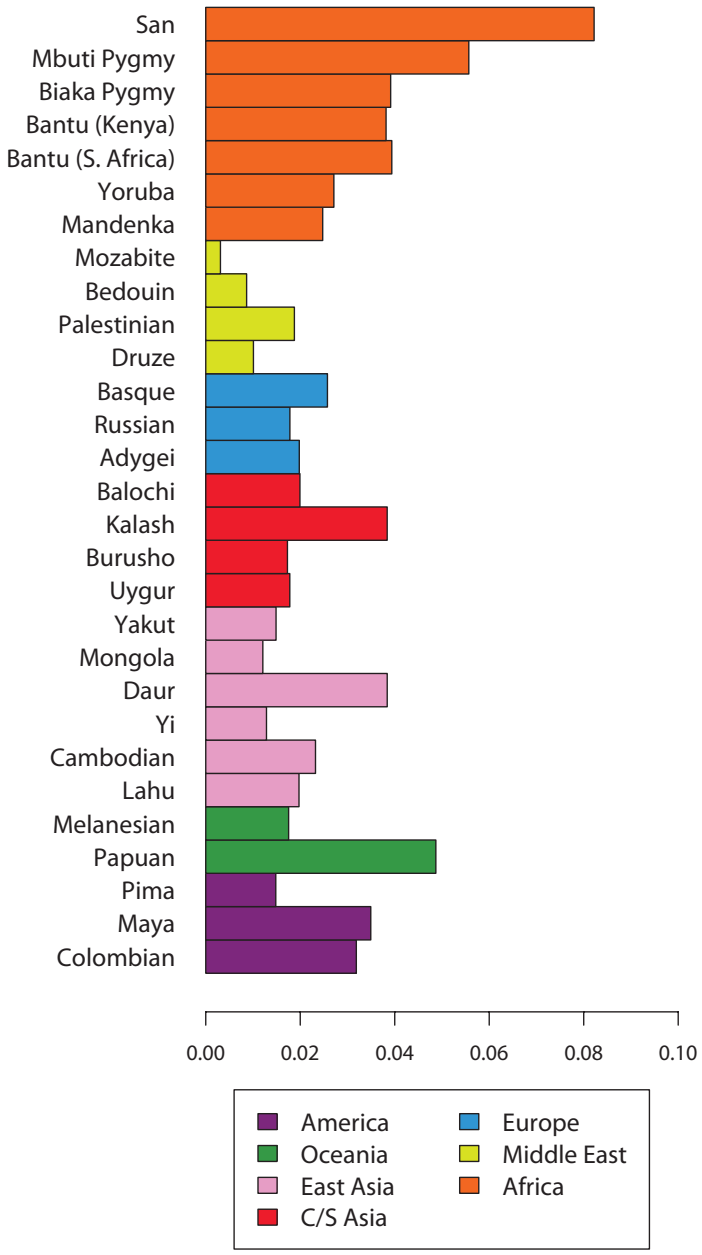

Figure S2.3: Difference in maximal imputation accuracy for two sets of SNPs (MAF>0.2 and MAF$\leq$0.2), based on data in Figure 2.9. Bars are colored by geographic locations of the populations.

Table S2.1: Imputation accuracy for inference of genotypes of untyped markers in the data of Jakobsson *et al.* (2008), based on any one or two or all three HapMap reference panels (with their original size). These values were used in the scatter plot of Figure 2.7. For each population, the highest imputation accuracy obtained among the seven possible reference panels is highlighted in bold.

	YRI	CEU	CHB+JPT	YRI/CEU	YRI/ CHB+JPT	CEU/ CHB+JPT	ALL
San	**0.8912**	0.8226	0.8205	0.8873	0.8873	0.8379	0.8879
Mbuti Pygmy	**0.9018**	0.8342	0.8224	0.8999	0.8994	0.8411	0.8988
Biaka Pygmy	0.9145	0.8559	0.8459	**0.9176**	0.9126	0.8629	0.9150
Bantu (Kenya)	0.9360	0.8752	0.8705	**0.9396**	0.9362	0.8875	0.9382
Bantu (S. Africa)	0.9322	0.8559	0.8536	**0.9329**	0.9322	0.8647	0.9325
Yoruba	**0.9457**	0.8667	0.8600	0.9448	0.9437	0.8790	0.9448
Mandenka	**0.9419**	0.8692	0.8650	0.9414	0.9412	0.8854	0.9408
Mozabite	0.9301	0.9180	0.9032	**0.9458**	0.9415	0.9226	0.9455
Bedouin	0.9279	0.9407	0.9215	0.9486	0.9421	0.9423	**0.9486**
Palestinian	0.9342	0.9502	0.9296	**0.9550**	0.9476	0.9507	0.9543
Druze	0.9300	0.9552	0.9341	0.9562	0.9500	0.9562	**0.9569**
Basque	0.9337	0.9577	0.9361	0.9570	0.9503	0.9576	**0.9579**
Russian	0.9338	0.9597	0.9389	0.9590	0.9524	**0.9600**	0.9599
Adygei	0.9315	0.9593	0.9372	0.9593	0.9512	**0.9605**	0.9600
Balochi	0.9249	0.9511	0.9337	**0.9527**	0.9473	0.9516	0.9524
Kalash	0.9165	0.9469	0.9287	0.9488	0.9425	**0.9492**	0.9477
Burusho	0.9301	0.9487	0.9342	0.9487	0.9465	0.9506	**0.9509**
Uygur	0.9270	0.9471	0.9427	0.9503	0.9487	**0.9537**	0.9534
Yakut	0.9249	0.9454	0.9466	0.9468	0.9505	**0.9516**	0.9513
Mongola	0.9302	0.9444	0.9517	0.9509	0.9545	0.9549	**0.9553**
Daur	0.9278	0.9434	0.9543	0.9493	0.9552	0.9565	**0.9571**
Yi	0.9268	0.9431	0.9522	0.9510	0.9533	0.9543	**0.9545**
Cambodian	0.9301	0.9413	0.9487	0.9455	0.9509	**0.9518**	0.9511
Lahu	0.9345	0.9480	**0.9599**	0.9531	0.9598	0.9588	0.9597
Melanesian	0.9207	0.9332	0.9454	0.9471	0.9477	0.9475	**0.9497**
Papuan	0.9212	0.9268	0.9399	0.9407	0.9436	0.9419	**0.9444**
Pima	0.9321	0.9487	0.9481	0.9540	0.9552	0.9542	**0.9554**
Maya	0.9305	0.9539	0.9495	0.9543	0.9558	**0.9588**	0.9582
Colombian	0.9305	0.9493	0.9398	0.9499	0.9507	0.9517	**0.9539**

Table S2.2: Squared correlation coefficient, r^2, between the genotypes imputed from the data of Jakobsson *et al.* (2008) and those directly measured in the data of Conrad *et al.* (2006) and Pemberton *et al.* (2008). These values were used in the scatter plot of Figure 2.8. For each population, the highest r^2 value obtained among the seven possible reference panels is highlighted in bold.

	YRI	CEU	CHB+JPT	YRI/CEU	YRI/ CHB+JPT	CEU/ CHB+JPT	ALL
San	**0.7633**	0.6116	0.6200	0.7443	0.7470	0.6416	0.7341
Mbuti Pygmy	0.7340	0.6299	0.6235	**0.7397**	0.7331	0.6570	0.7346
Biaka Pygmy	0.7804	0.6708	0.6397	**0.7882**	0.7716	0.6785	0.7795
Bantu (Kenya)	0.8611	0.7178	0.7212	**0.8726**	0.8553	0.7387	0.8672
Bantu (S. Africa)	0.8452	0.6825	0.6833	**0.8510**	0.8454	0.6842	0.8492
Yoruba	**0.8999**	0.7155	0.7049	0.8951	0.8924	0.7350	0.8957
Mandenka	**0.8744**	0.7087	0.6978	0.8670	0.8731	0.7327	0.8671
Mozabite	0.8541	0.8253	0.7833	0.8959	0.8911	0.8339	**0.8973**
Bedouin	0.8574	0.8830	0.8296	**0.9067**	0.8871	0.8843	0.9062
Palestinian	0.8678	0.9015	0.8526	**0.9157**	0.8984	0.9002	0.9095
Druze	0.8525	0.9107	0.8588	0.9123	0.8943	0.9156	**0.9161**
Basque	0.8686	0.9240	0.8869	0.9214	0.9023	0.9234	**0.9262**
Russian	0.8682	**0.9310**	0.8689	0.9241	0.9093	0.9292	0.9291
Adygei	0.8620	0.9277	0.8749	0.9296	0.9006	**0.9307**	0.9269
Balochi	0.8614	0.9099	0.8724	0.9175	0.8943	**0.9185**	0.9155
Kalash	0.8585	0.9058	0.8562	0.9116	0.8931	**0.9135**	0.9069
Burusho	0.8896	0.9082	0.8699	0.9067	0.9059	0.9127	**0.9161**
Uygur	0.8675	0.9114	0.8986	0.9175	0.9188	**0.9300**	0.9282
Yakut	0.8633	0.9036	0.9102	0.9078	0.9205	**0.9239**	0.9225
Mongola	0.8803	0.9066	0.9212	0.9180	**0.9265**	0.9236	0.9257
Daur	0.8612	0.8968	0.9289	0.9147	0.9286	0.9300	**0.9305**
Yi	0.8665	0.8947	0.9127	0.9069	0.9181	0.9199	**0.9212**
Cambodian	0.8752	0.8858	0.9102	0.8962	0.9156	**0.9165**	0.9114
Lahu	0.8832	0.8978	0.9332	0.9098	**0.9335**	0.9291	0.9323
Melanesian	0.8504	0.8597	0.8986	0.9002	0.9035	0.9042	**0.9057**
Papuan	0.8581	0.8638	0.8796	0.8762	0.8855	**0.8884**	0.8839
Pima	0.9271	0.9486	0.9423	0.9604	0.9509	0.9572	**0.9618**
Maya	0.8738	0.9210	0.9146	0.9051	0.9182	**0.9243**	0.9196
Colombian	0.8858	0.9309	0.9101	0.9209	0.9302	0.9296	**0.9331**

Table S2.3: Summary statistics for minor allele frequencies of 513 SNP loci in the data of Conrad *et al.* (2006) and Pemberton *et al.* (2008). The statistics reported here correspond to those of the marker sets that yielded the imputation accuracy plotted in Figure 2.9.

	All		MAF<0.2			MAF≥0.2		
	Mean	Standard deviation	Number of SNPs	Mean	Standard deviation	Number of SNPs	Mean	Standard deviation
San	0.1861	0.1639	294	0.0620	0.0672	219	0.3528	0.0912
Mbuti Pygmy	0.1988	0.1521	271	0.0751	0.0616	242	0.3372	0.0921
Biaka Pygmy	0.2270	0.1570	252	0.0886	0.0631	261	0.3607	0.0905
Bantu (Kenya)	0.2474	0.1487	206	0.0925	0.0625	307	0.3513	0.0859
Bantu (S. Africa)	0.2270	0.1445	253	0.1000	0.0707	260	0.3506	0.0731
Yoruba	0.2419	0.1444	213	0.0985	0.0564	300	0.3437	0.0916
Mandenka	0.2370	0.1453	227	0.0998	0.0581	286	0.3460	0.0914
Mozabite	0.2670	0.1259	168	0.1196	0.0521	345	0.3387	0.0807
Bedouin	0.2564	0.1284	179	0.1181	0.0599	334	0.3305	0.0875
Palestinian	0.2539	0.1378	204	0.1159	0.0649	309	0.3450	0.0886
Druze	0.2468	0.1431	200	0.1002	0.0603	313	0.3404	0.0935
Basque	0.2299	0.1503	255	0.1008	0.0602	258	0.3575	0.0923
Russian	0.2235	0.1400	223	0.0935	0.0549	290	0.3235	0.0966
Adygei	0.2383	0.1446	231	0.1016	0.0527	282	0.3502	0.0888
Balochi	0.2459	0.1408	173	0.0877	0.0549	340	0.3264	0.0955
Kalash	0.2490	0.1460	202	0.0959	0.0659	311	0.3484	0.0850
Burusho	0.2628	0.1521	188	0.0882	0.0573	325	0.3638	0.0822
Uygur	0.2636	0.1410	167	0.0961	0.0512	346	0.3444	0.0900
Yakut	0.2383	0.1484	182	0.0692	0.0507	331	0.3313	0.0913
Mongola	0.2507	0.1515	198	0.0903	0.0631	315	0.3515	0.0923
Daur	0.2303	0.1473	206	0.0823	0.0604	307	0.3297	0.0959
Yi	0.2481	0.1541	177	0.0715	0.0585	336	0.3412	0.0967
Cambodian	0.2510	0.1528	216	0.0943	0.0738	297	0.3649	0.0741
Lahu	0.2223	0.1603	261	0.0843	0.0745	252	0.3652	0.0798
Melanesian	0.2155	0.1761	261	0.0589	0.0646	252	0.3777	0.0838
Papuan	0.2113	0.1626	240	0.0590	0.0622	273	0.3453	0.0889
Pima	0.2047	0.1828	237	0.0267	0.0445	276	0.3576	0.0987
Maya	0.2142	0.1633	256	0.0703	0.0656	257	0.3575	0.0878
Colombian	0.2157	0.1617	227	0.0607	0.0599	286	0.3387	0.0991

CHAPTER III

The Relationship between Imputation Error and Statistical Power in Genetic Association Studies in Diverse Populations

The genotype imputation strategy for case-control genetic association studies provides an economical way of assessing many more genetic markers for disease association than have actually been measured in any particular association study (Li *et al.*, 2006; Nicolae, 2006; Marchini *et al.*, 2007; Servin & Stephens, 2007; Browning, 2008). In this approach, case and control individuals are first genotyped for markers densely spread across the human genome. The genotypes obtained are then combined with high-resolution genotypic data from genomic databases to impute the genotypic status of study individuals at markers investigated in the database but not in the study sample. This imputation relies on the principle that two haplotypes identical in genotype at nearby SNP markers are likely to share intervening chromosomal stretches identically by descent. Thus, if a haplotype in a densely genotyped database sample is identical to a haplotype in a more sparsely genotyped study sample for markers that overlap between the study and the database, then the study haplotype can be imputed with high resolution by copying the haplotype from the database.

Partly because they dramatically increase the number of markers that can be di-

rectly tested for association compared to earlier tag-SNP designs, methods relying on genotype imputation have proven effective for identifying high-risk disease-associated genetic variants (Scott *et al.*, 2007; Wellcome Trust Case Control Consortium, 2007; Barrett *et al.*, 2008; Zeggini *et al.*, 2008; Willer *et al.*, 2008). However, the imputation strategy utilizes in its association tests estimated genotypes that are not known with certainty, and errors in imputed genotypes might potentially compromise the power of an imputation-based association test. For example, at a biallelic marker, consider a disease-susceptibility allele of small effect that has true frequency 0.3 in cases and 0.2 in controls. If the probability that imputation recovers the true allele is 0.9, then the frequency of the disease allele among *imputed* genotypes will be $(0.3)(0.9) + (0.7)(0.1) = 0.34$ in cases and $(0.2)(0.9) + (0.8)(0.1) = 0.26$ in controls. Imputation error converts an allele frequency difference of $0.3 - 0.2 = 0.1$ between cases and controls into a smaller difference of $0.34 - 0.26 = 0.08$. As a result, for the imputed genotypes, a larger sample size might be required for determining that allele frequencies differ between cases and controls compared to the sample size that would be required if the true genotypes were known.

Although recent studies have found that imputation error rates are generally low (Yu & Schaid, 2007; Guan & Stephens, 2008; Pei *et al.*, 2008; Zhao *et al.*, 2008; Nothnagel *et al.*, 2009), it is possible that even low error rates could have considerable effects on downstream analyses. How does the error inherent in genotype imputation reduce the power of an association study when alleles at the true disease SNP are imputed rather than known? An answer to this question is important to the design and interpretation of imputation-based association studies. Relating imputation error and power would assist in calculating sample sizes required for detecting disease variants at loci whose genotypes are imputed, and for determining if imputation studies in particular populations are likely to be underpowered. Additionally, a relationship between imputation error and power would aid in developing resources for genomic

studies. For example, use of such a relationship could assist in identifying populations in whom existing resources produce high error rates that limit the potential for practical mapping of risk variants with imputation strategies.

The problem of connecting imputation error to power is similar to a corresponding problem in the context of tag SNPs. In the imputation context, the loss of information due to imputation error at a disease-susceptibility locus can obscure the association between the locus and disease. In the tag-SNP context, the loss of information due to use of a tag SNP rather than the true disease SNP has an analogous effect. In both situations, missing information about the correct genotypes at the true disease-susceptibility locus contributes to a loss of power to detect disease association.

For the tag-SNP context, consider two loci, a SNP causally associated with disease and a nearby tag SNP. If the r^2 correlation statistic for linkage disequilibrium (LD) between the tag SNP and the disease SNP is equal to c, then a chi-squared test statistic for disease association at the true disease SNP in a case-control sample of total size N has approximately the same asymptotic distribution under the alternative hypothesis of disease association as the corresponding chi-squared statistic at the tag SNP in a case-control sample of size N/c (Pritchard & Przeworski, 2001). Thus, the "sample size inflation factor" required in using the tag SNP in an association study rather than the true disease SNP is $\sim 1/c$.

Motivated by this result, multiple versions of an r^2 correlation statistic between the imputed genotypes at a SNP and the true genotypes have been proposed (de Bakker *et al.*, 2008; Browning & Browning, 2009; Huang *et al.*, 2009a). Such statistics, which are sometimes used to identify markers imputed with high accuracy in imputation-based GWA studies (Scott *et al.*, 2007; Lettre *et al.*, 2008), have been viewed as conceptually analogous to the r^2 statistic for LD between a tag SNP and a disease SNP, but have not been shown to be mathematically equivalent to it. In the imputation context for a biallelic SNP with alleles A and B, the correlation between

true and imputed genotypes is a function of a 3×3 table, in which each of three possible true genotypes (AA, AB, BB) has one of three possible imputations. In the tag-SNP context, however, if the disease SNP has alleles A and B and the tag SNP has alleles C and D, then the corresponding table is 2×2, containing entries for the counts of the four possible haplotypes (AC, AD, BC, BD). Although the close analogy between the tag SNP and imputation contexts suggests that the relationship between imputation error and power is similar to that observed between power and LD with a tag SNP, at present the connection between imputation r^2 statistics and power remains informal.

Here, to investigate the mathematical relationship between imputation error and power, we adapt a method developed for evaluating the relationship between *genotyping* error and power (Gordon *et al.*, 2002; Kang *et al.*, 2004). Our approach does not use an r^2 statistic, and unlike in the tag-SNP context, in which the inflation factor depends only on the LD between the tag and disease SNPs, the corresponding inflation factor in the imputation context is a function of nine parameters. Consider two 2×3 chi-squared tests of association, examining the relationship between the three possible genotypes of a biallelic marker and case-control status. The first test uses the true genotypes of the marker, whereas the second test uses genotypes measured with the possibility of imputation error. Suppose that k is the ratio of the number of controls to the number of cases. Denote by $\mathrm{MAF}_{controls}$ the frequency of the minor allele in controls, and by MAF_{cases} the frequency of this same allele in cases. Thus, $0 \leq \mathrm{MAF}_{controls} \leq 1/2$ and $0 \leq \mathrm{MAF}_{cases} \leq 1$. We label the minor allele in controls by A, the major allele in controls by B, genotype AA by 1, AB by 2, and BB by 3. For $i, j \in \{1, 2, 3\}$, we let ϵ_{ij} be the probability that genotype i is imputed as genotype j. Because $\sum_{j=1}^{3} \epsilon_{ij} = 1$ for each i, only six error parameters must be considered: ϵ_{12}, ϵ_{13}, ϵ_{21}, ϵ_{23}, ϵ_{31}, and ϵ_{32}.

Gordon *et al.* (2002) and Kang *et al.* (2004) determined the relationship between

the two 2×3 chi-squared test statistics at a locus, showing that the test statistic for association between true genotype and disease in a sample of size N has the same asymptotic distribution as the test statistic for association between imputed genotype and disease in a sample of size Nf, where $f \geq 1$ is a rational function of ϵ_{12}, ϵ_{13}, ϵ_{21}, ϵ_{23}, ϵ_{31}, ϵ_{32}, k, MAF_{cases} and $\mathrm{MAF}_{controls}$ that represents the sample size inflation factor. Thus, if a sample of size at least N is required for achieving a specified level of power when genotype is measured without error, then a sample of size at least Nf is required for achieving the same power when genotype is imputed with error. We use a special case of the formula for f, assuming $k = 1$, so that a study has equally many cases and controls; we also assume Hardy-Weinberg proportions are satisfied separately in cases and controls. With these assumptions, the sample size inflation factor due to imputation error can be written $f = g/g^*$, defining g and g^* as in eqs. 1 and A.1 of Kang *et al.* (2004), and matching our notation to that of Kang *et al.* (2004) with the substitutions $P_{01} = \mathrm{MAF}_{cases}^2$, $P_{02} = 2\mathrm{MAF}_{cases}(1 - \mathrm{MAF}_{cases})$, $P_{03} = (1 - \mathrm{MAF}_{cases})^2$, $P_{11} = \mathrm{MAF}_{controls}^2$, $P_{12} = 2\mathrm{MAF}_{controls}(1 - \mathrm{MAF}_{controls})$, and $P_{13} = (1 - \mathrm{MAF}_{controls})^2$.

To evaluate the sample size inflation factor f at levels of imputation error appropriate for typical association studies, we first estimated the six error parameters using genotypes of 426 individuals in 29 diverse populations. Employing reference panels of phased haplotypes based on ∼2,000,000 SNPs in 210 HapMap Phase II individuals together with a worldwide study of ∼500,000 SNPs (Jakobsson *et al.*, 2008), we imputed individual genotypes at markers that were included in the reference data but not in the worldwide study. For each population, we repeated the imputations underlying Figure 7 of Huang *et al.* (2009a), using the same procedure as was used by Huang *et al.* (2009a), to obtain an imputed data set of 513 markers. This set consisted of probabilistic imputations relying on the subset of reference individuals that in the work of Huang *et al.* (2009a) produced the highest imputation accuracy

for that population, among seven choices. The genotypes of Pemberton *et al.* (2008), which update those reported by Conrad *et al.* (2006), were treated as true genotypes of the 513 markers for measurement of ϵ_{ij}. For each population, at each marker, the minor and major alleles were determined only using the "true" genotype data from that population. If each allele had frequency 50%, then the minor allele was assigned at random.

Treating the 426 individuals as unaffected, we classified 218,345 true genotypes (426×513, excluding missing data) by category, and separately for each population, we estimated ϵ_{12}, ϵ_{13}, ϵ_{21}, ϵ_{23}, ϵ_{31}, and ϵ_{32}. Each true genotype was categorized as follows: 1—minor allele homozygote; 2—heterozygote; 3—major allele homozygote. Considering all true genotypes in a population at all 513 markers, denote the number of true genotypes of types 1, 2, and 3 by n_1, n_2, and n_3, respectively. For each population, n_1, the smallest of the three quantities, was at least 70, so that at least 70 true genotypes were used in estimating each error parameter. For n_1, n_2, and n_3, the medians across populations were 411, 1967, and 3679, respectively.

To incorporate the uncertainty inherent in imputing a genotype, posterior probabilities of imputing types 1, 2, and 3 were obtained. Considering the n_i genotypes of type i, denote the posterior probability that genotype ℓ was imputed to have type j by $q_{ij\ell}$. For each $i, j \in \{1, 2, 3\}$, $i \neq j$, we computed ϵ_{ij} for the population as $\sum_{\ell=1}^{n_i} q_{ij\ell}/n_i$. The "overall imputation error rate," a weighted average of the ϵ_{ij} that evaluates the total fraction of alleles imputed incorrectly, was calculated as $[(\frac{1}{2}\epsilon_{12} + \epsilon_{13})n_1 + (\frac{1}{2}\epsilon_{21} + \frac{1}{2}\epsilon_{23})n_2 + (\epsilon_{31} + \frac{1}{2}\epsilon_{32})n_3]/(n_1 + n_2 + n_3)$.

For each population, Figure 3.1 displays the estimated values of ϵ_{ij}. In most populations, the highest imputation error rate is ϵ_{12}, indicating that conditional on true genotype the highest-probability error is misclassification of a minor allele homozygote as a heterozygote. The next highest error rate is usually ϵ_{13} or ϵ_{23}, reflecting misclassification probabilities for minor allele homozygotes or heterozygotes, respec-

tively, as major allele homozygotes. Misclassification probabilities for major allele homozygotes or heterozygotes as minor allele homozygotes (ϵ_{31}, ϵ_{21}) are generally low.

Treating the estimated values of ϵ_{ij} as parametric values, for each population, we evaluated the sample size inflation factor f for various choices of the unknown MAF_{cases} and $\text{MAF}_{controls}$. Because the difference $\delta = \text{MAF}_{cases} - \text{MAF}_{controls}$ can be viewed as a measure of the magnitude of the association at a disease locus, we reparametrized f in terms of δ and $\text{MAF}_{controls}$. Thus, using observed levels of imputation error, we examined the properties of f across the range of possible frequencies for the disease allele in cases and controls (Figure 3.2). For most choices of the parameter values in most populations, the inflation factor f lies between 1.1 and 1.6. For most African populations, consistent with their higher imputation error rates, f is considerably greater than in other populations, ranging from 1.3 to 2.5 for most choices of the parameter values. The inflation factor is especially high in the San and Mbuti Pygmy populations, in which nearly all choices examined for δ and $\text{MAF}_{controls}$ produce $f \gtrsim 1.7$. Disease alleles are difficult to detect when $|\delta|$ is small, and Figure 3.2 demonstrates that for several populations, the sample size inflation factor is greatest for small $|\delta|$, particularly when the disease locus has a low minor allele frequency of $\text{MAF}_{controls} = 0.05$.

Because the parameters MAF_{cases} and $\text{MAF}_{controls}$ are unknown in actual association studies, for each population, conditional on the imputation error parameters ϵ_{ij}, we examined the minimal and maximal values of the sample size inflation factor f across the range of possible values for MAF_{cases} and $\text{MAF}_{controls}$ (Figure 3.3). For most non-African populations, considering the range of possible values for the minor allele frequency in controls, the minimal f is typically in the range 1.1-1.2 and the maximum is typically in the range 1.2-1.6, indicating that the extra sample size required for maintaining power is usually at least 10-20% and at most 20-60%. The

maximal f is generally greater for low values of $\text{MAF}_{controls}$.

Examining the minimal and maximal sample size inflation factor across the range of disease allele frequencies (Figure 3.3), the values are greatest in populations with the highest imputation error rates (Figure 3.1). Figure 3.4 quantifies this observation, illustrating the relationships with overall imputation error rate of the minimal and maximal values of f. A linear regression of the minimal sample size inflation factor on overall imputation error rate when $\text{MAF}_{controls}$ is fixed at 0.3, forced through the point at which no imputation errors occur and therefore no sample size inflation occurs, provides a close fit for most populations, with the exceptions of the San and Mbuti Pygmy populations. The slope for this regression is 6.911, and the corresponding regression for the maximal sample size inflation factor has a slope of 10.177. Excluding the San and Mbuti Pygmy populations, the slopes of the regressions for the minimal and maximal sample size inflation factors decrease to 6.203 and 8.836, respectively (Figure S3.1). More generally, the regression slopes generally lie between 5 and 13 when fixing $\text{MAF}_{controls}$ at various values across its range, either including or excluding the San and Mbuti Pygmy populations (Figures S3.1, S3.2). These values have the interpretation that each 1% increase in overall imputation error rate translates to an increase of ∼5-13% in the sample size required for maintaining power.

Our results have important implications for imputation studies. In the tag-SNP setting, for small x, a high LD level of $r^2 = 1 - x$ produces a relatively small sample size inflation factor of $1/(1-x) \approx 1+x$, so that each 1% loss in the r^2 measure of LD leads to a ∼1% gain in the required sample size. In the imputation setting, however, imputation accuracy of $1-x$ produces a typical inflation factor of $\sim 1+5x$ to $\sim 1+13x$, so that each 1% loss in imputation accuracy leads to a ∼5-13% increase in the required sample size. As a result, even low levels of imputation error can have sizeable consequences. For example, measures that aim to assess genomic coverage for imputation methods might need to require stringent levels of imputation error in evaluating the

proportion of the genome that is suited to imputation-based association mapping. Studies that aim to confirm associations at imputed markers in populations with lower imputation accuracy might inherently be disadvantaged for success in replication studies. In these various settings, careful assessment of appropriate sample sizes in power calculations will be essential for progress in imputation-based disease-gene identification. One key observation is that imputation error produces the greatest sample-size inflation for markers with low minor allele frequency ($\text{MAF}_{controls} \leq 0.1$), and for such markers the sample-size inflation for each 1% imputation error can be as high as ∼15-35% (Figures S3.1, S3.2). As GWA efforts begin to focus on the impact of rare alleles on complex diseases, the potentially serious effects of imputation error for detecting such alleles will be a central consideration for forthcoming studies. For such studies, it will be informative to examine values of the imputation error parameters ϵ_{ij} evaluated specifically from rare alleles.

We note that the linear dependence of the minimal and maximal sample size inflation factor on overall imputation error rate, as illustrated in Figure 3.4, is only approximate. This approximate linear relationship arises because the overall imputation error rate is a composite parameter dependent on the six underlying ϵ_{ij} parameters, each of which affects the inflation factor in an approximately linear manner. Based on a first-order Taylor series expansion for f, for each i and j, Kang *et al.* (2004) derived cost functions C_{ij} so that if all error parameters except ϵ_{ij} are set to zero and ϵ_{ij} is small, then the sample size inflation factor is approximately $1 + C_{ij}\epsilon_{ij}$. These linear approximations accurately reflect the sample size inflation factor in most populations except at the lowest values of MAF_{cases} and $\text{MAF}_{controls}$ (results not shown), and suggest that in general, the greatest cost is incurred from errors in imputing minor allele homozygotes as major allele homozygotes (Figure 3.5). It is noteworthy that the linear regressions in Figure 3.4 provide the poorest underestimates in the San population, for which the parameter ϵ_{13} for the most costly type of error was

high, and for which the pattern of errors differed somewhat from the corresponding patterns in the other populations (Figure 3.1).

While an increased sample size provides one approach to maintaining power in an imputation-based study, an alternative strategy is to instead decrease imputation error. Reductions in imputation error can be achieved through a combination of algorithmic advances and optimal choices of imputation algorithms (Pei *et al.*, 2008; Nothnagel *et al.*, 2009), improvements in usage of existing reference panels (Huang *et al.*, 2009a; Howie *et al.*, 2009), and expanded marker density and sample inclusion in these panels (Browning & Browning, 2009; Becker *et al.*, 2009). A fourth approach involves incorporating information on relatives of study subjects to improve phase estimates at measured markers; although this approach will not eliminate errors owing to incorrect imputation conditional on correctly estimated phase, it will reduce imputation errors that arise from incorrect phase estimation.

For populations with relatively little imputation error, in which large samples are easily obtained, the required sample size increase produced by imputation error might not pose a significant obstacle for GWA studies. In other populations in which subject recruitment is difficult and the sample size inflation required for maintaining power is extreme, reduction of imputation error might be more feasible than an increase in sample size. As GWA studies begin diversifying to incorporate additional populations beyond the populations of European origin that have been typical of most investigations to date (Cooper *et al.*, 2008), it will be important to evaluate the relative merits of the various approaches for overcoming the consequences of imputation error to improve the potential of imputation-based association studies.

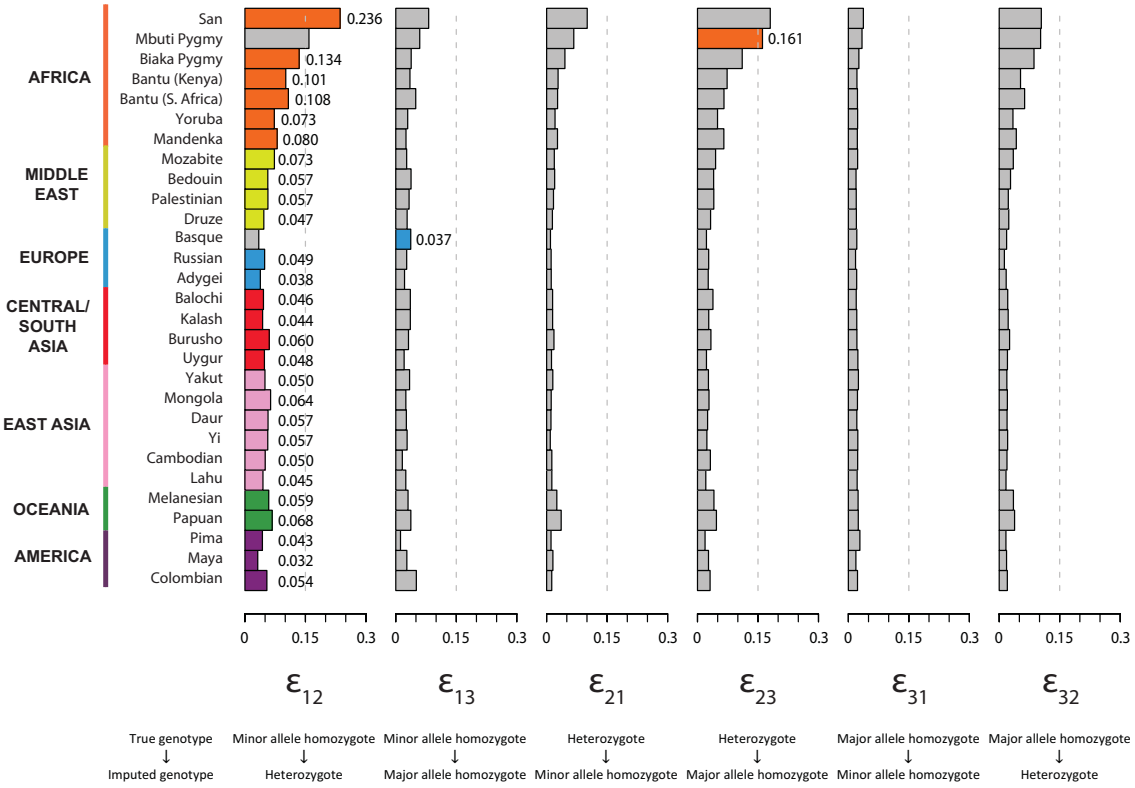

Figure 3.1: Genotype misclassification rates at imputed loci, in each of 29 populations. Each bar plot presents a particular error rate ϵ_{ij}, where ϵ_{ij} represents the probability that genotype i is imputed as genotype j (1—minor allele homozygote, 2—heterozygote, 3—major allele homozygote). For each population, the greatest of the six error rates is shown in color, with a color characteristic of the geographic region of the population. For convenience in interpreting the figure, the vertical dashed line indicates 15% error. The values plotted in the figure appear together with the overall imputation error rate in Table S3.1.

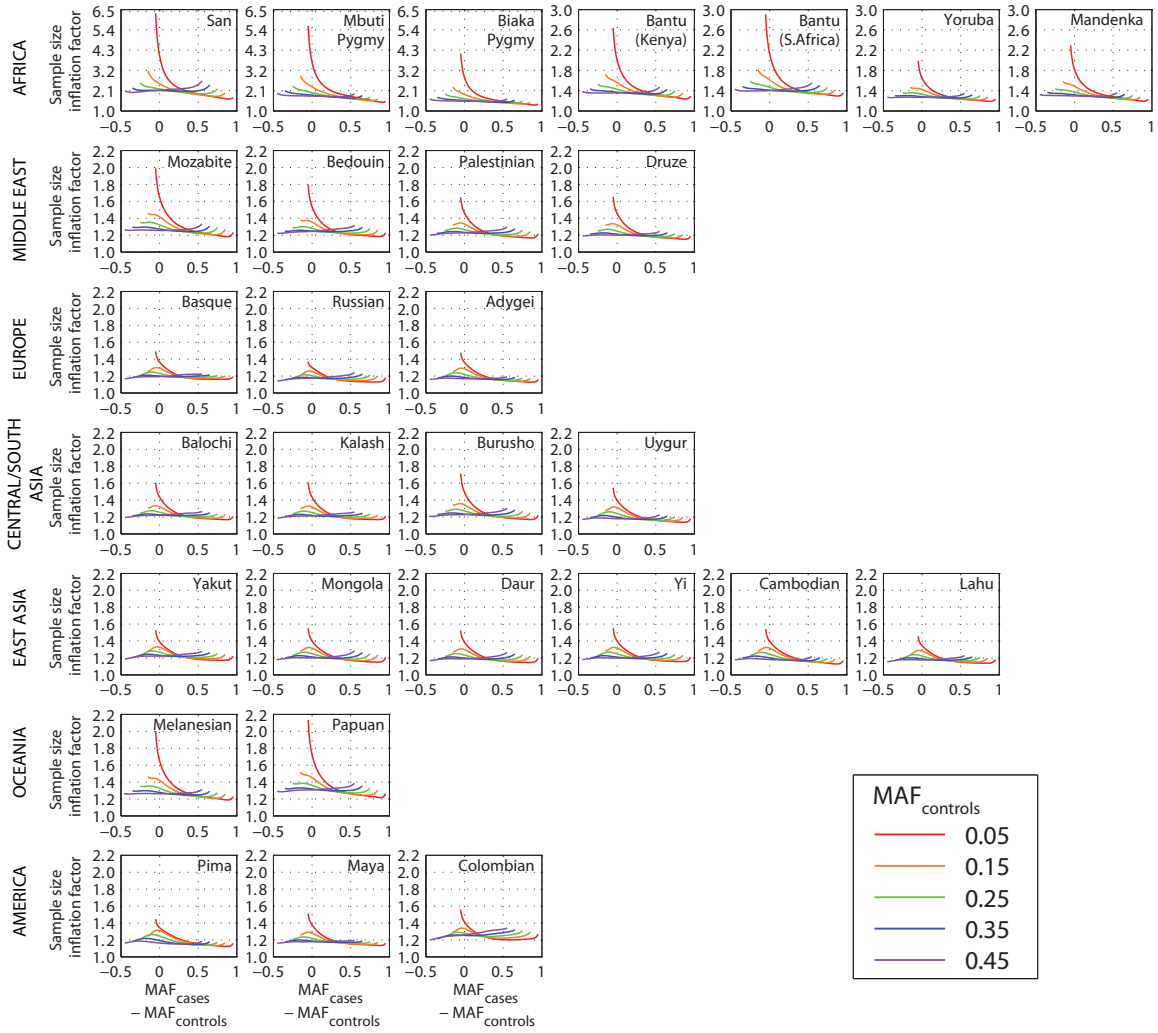

Figure 3.2: Sample size inflation factor f required for maintaining statistical power at imputed loci, as a function of the true difference in the frequency of the minor allele between cases and controls. Each plot utilizes the estimated imputation error rates in Figure 3.1 for a specific population. For each population, the inflation factor is plotted for five choices of the true minor allele frequency in controls (0.05, 0.15, 0.25, 0.35, and 0.45). Note that $\text{MAF}_{controls}$ ranges from 0 to 0.5, whereas MAF_{cases}, representing the frequency in cases of the minor allele in controls, ranges from 0 to 1. We used a step size of 0.001 for MAF_{cases} and disregarded points with $\text{MAF}_{cases} = \text{MAF}_{controls}$.

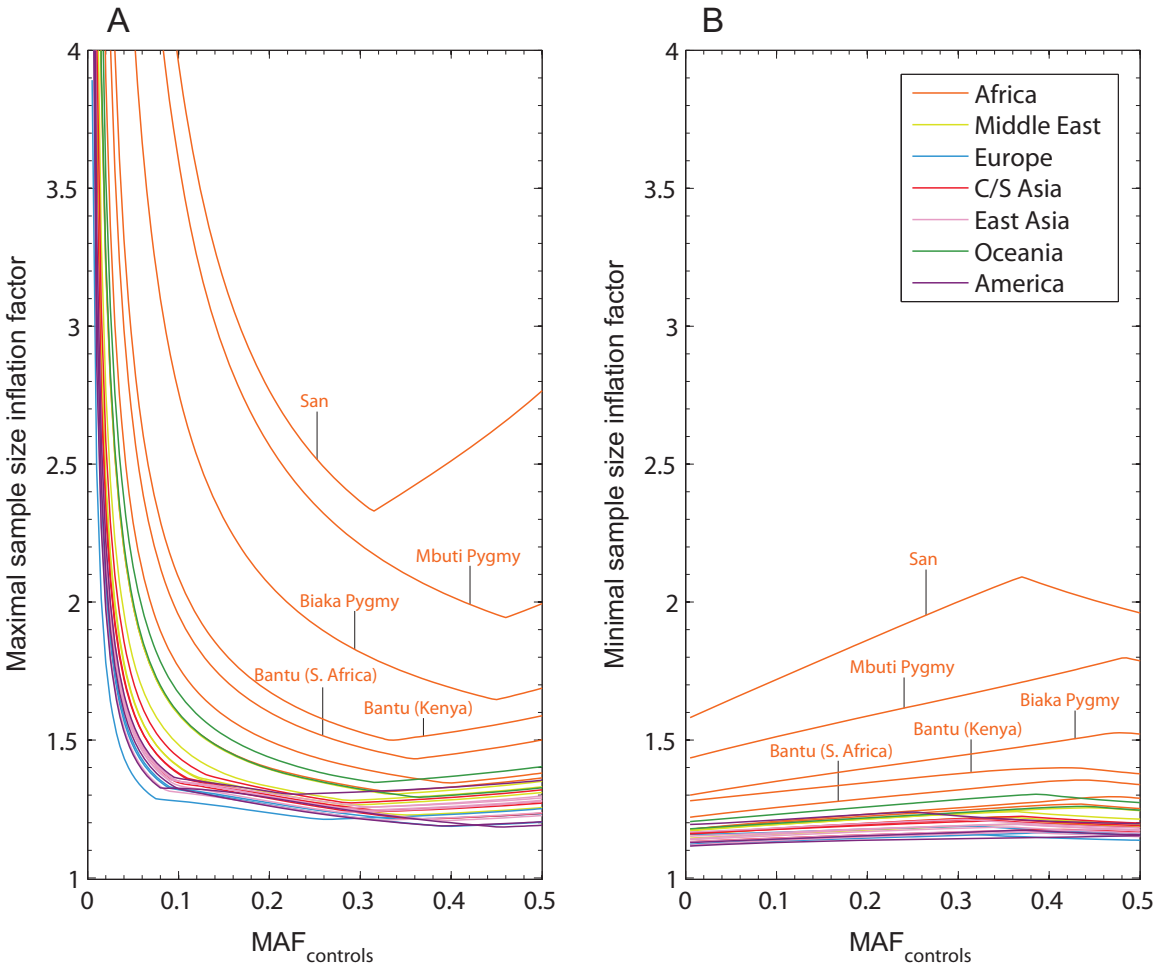

Figure 3.3: Maximal and minimal sample size inflation factor at imputed loci as functions of the true minor allele frequency in controls, in each of 29 populations. For each value of $\text{MAF}_{controls}$ from 0.005 to 0.5 with a step size of 0.005, the value plotted is the maximal or minimal value of the inflation factor f, considering choices of MAF_{cases} ranging from 0 to 1 with a step size of 0.001 ($\text{MAF}_{cases} \neq \text{MAF}_{controls}$). Graphs for individual populations are color-coded by geographic region. (A) Maximal sample size inflation factor. (B) Minimal sample size inflation factor.

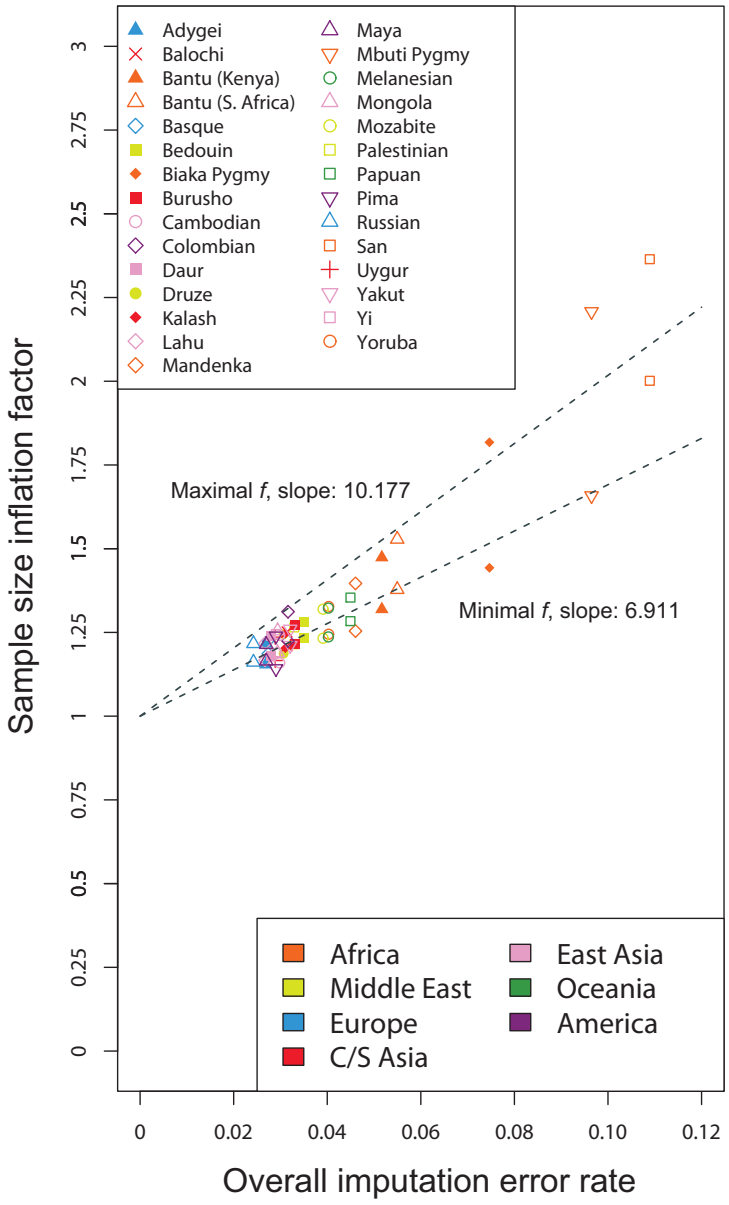

Figure 3.4: Maximal and minimal sample size inflation factor as functions of the overall imputation error rate, for an imputed disease locus with true minor allele frequency 0.3 in controls. Populations are color-coded by geographic region, and two data points appear for each population, a maximum and a minimum. Best-fit linear regression lines for the maxima and minima, forced through the point (0,1), indicate the increase in the inflation factor with increasing imputation error rate. For example, the lines indicate that in most populations, at $\mathrm{MAF}_{controls} = 0.3$, imputation error rates of 2-6% correspond to sample size inflation factors of ∼14-53%, and each additional increase of 1% in imputation error corresponds to an increase of ∼7-10% in the inflation factor.

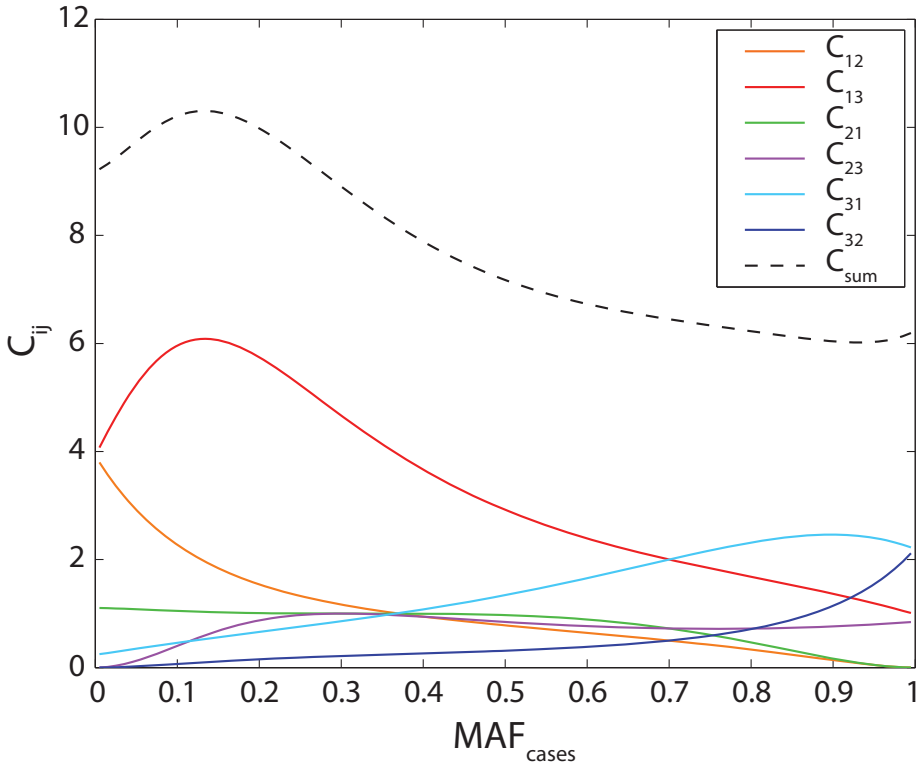

Figure 3.5: Cost coefficients as functions of MAF_{cases} for the fixed value $\text{MAF}_{controls} = 0.3$. The coefficient C_{ij} provides an approximation to the relative magnitude of the sample size inflation due to the error parameter ϵ_{ij}. Thus, a small increase of x in the imputation error parameter ϵ_{ij} adds approximately $C_{ij}x$ to the sample size inflation factor. The sum of the six cost coefficients, C_{sum}, has the interpretation that $C_{sum}x$ is added to the sample size inflation factor when all six of the ϵ_{ij} are simultaneously set to x. Each of the cost coefficients was evaluated for values of MAF_{cases} from 0.005 to 0.995 at intervals of 0.01.

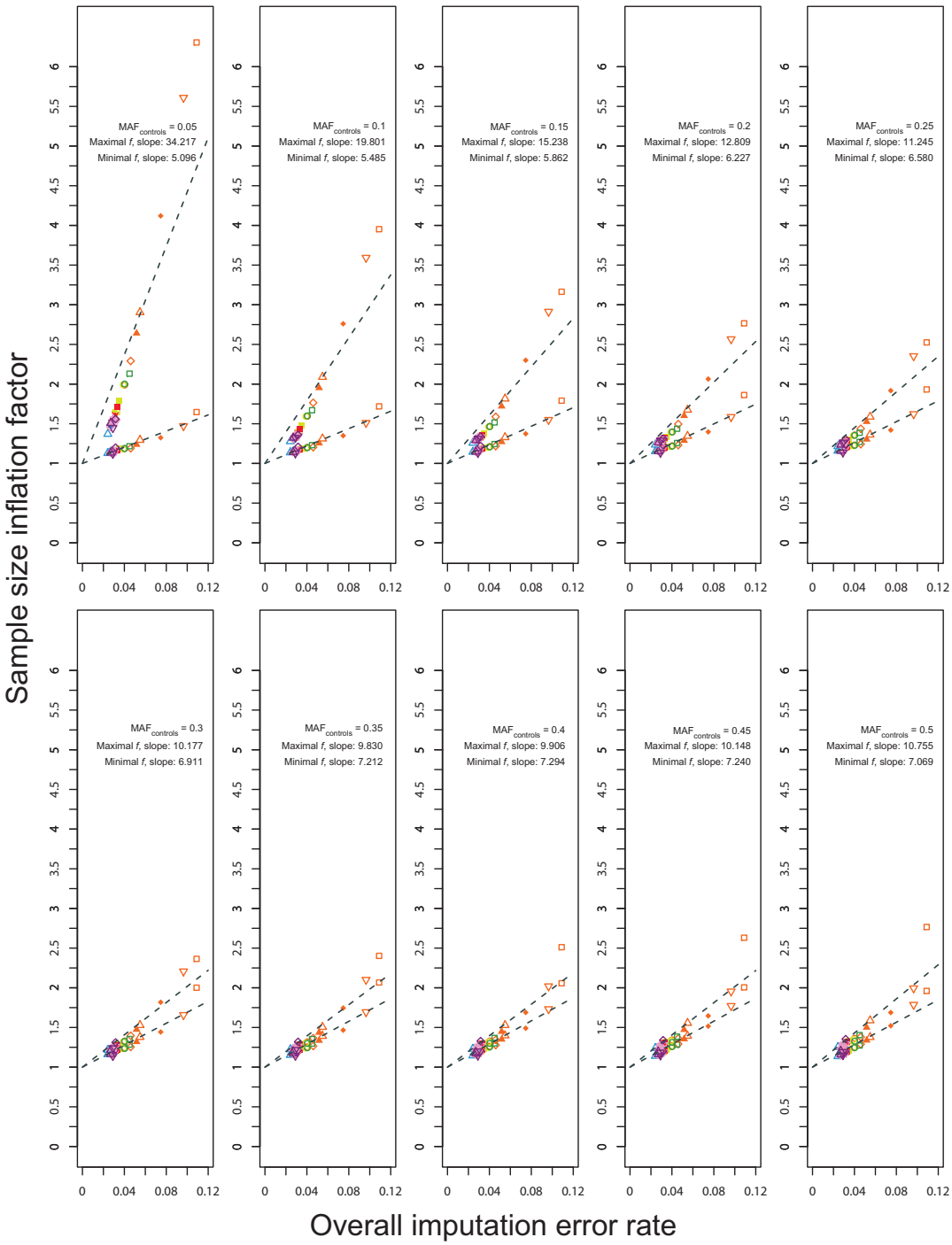

Figure S3.1: Maximal and minimal sample size inflation factor as functions of the overall imputation error rate, for an imputed disease locus with true minor allele frequency fixed in controls, excluding the San and Mbuti Pygmy populations. Each plot has $\text{MAF}_{controls}$ fixed at a different value. Population symbols are the same as in Figure 3.4, and two data points appear for each population, a maximum and a minimum. Best-fit linear regression lines for the maxima and minima, forced through the point (0,1), indicate the increase in the inflation factor with increasing imputation error rate.

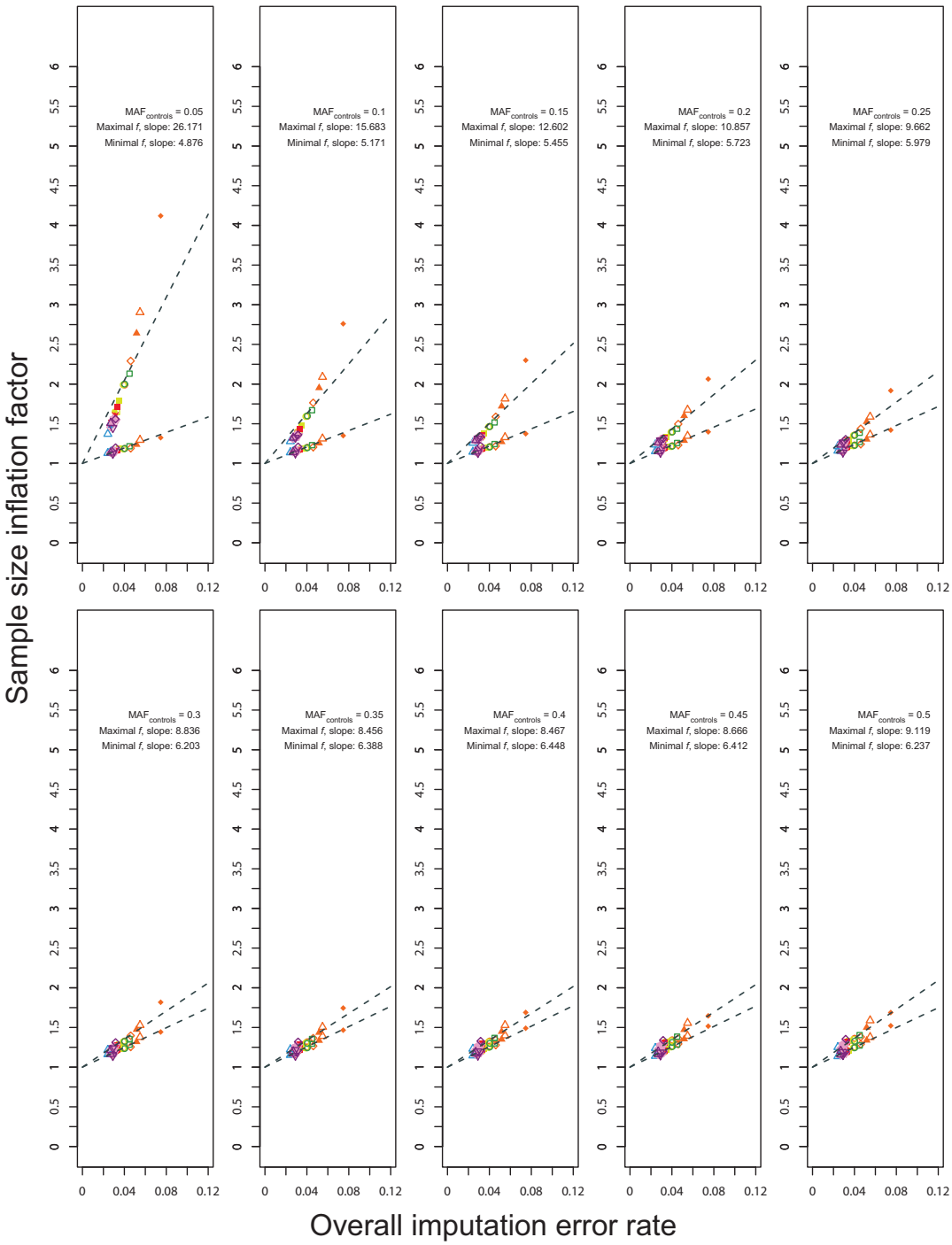

Figure S3.2: Maximal and minimal sample size inflation factor as functions of the overall imputation error rate, for an imputed disease locus with true minor allele frequency fixed in controls, considering all 29 populations. Each plot has $\text{MAF}_{controls}$ fixed at a different value. Population symbols are the same as in Figure 3.4, and two data points appear for each population, a maximum and a minimum. Best-fit linear regression lines for the maxima and minima, forced through the point (0,1), indicate the increase in the inflation factor with increasing imputation error rate. The plot for $\text{MAF}_{controls} = 0.3$ also appears in Figure 3.4.

Table S3.1: Genotype misclassification error rates ϵ_{ij} in each of 29 populations. Each column shows a particular error rate ϵ_{ij}, where ϵ_{ij} represents the probability that genotype i is imputed as genotype j (1—minor allele homozygote, 2—heterozygote, 3—major allele homozygote). For each population, the greatest of the six error rates is shown in bold. The values of ϵ_{ij} are identical to those plotted in Figure 3.1. The final column represents the overall imputation error rate, a weighted sum of the six ϵ_{ij} with the weights determined by the frequencies of the three categories of genotype.

Population	ϵ_{12}	ϵ_{13}	ϵ_{21}	ϵ_{23}	ϵ_{31}	ϵ_{32}	Overall imputation error rate
San	**0.236**	0.082	0.101	0.180	0.037	0.104	0.109
Mbuti Pygmy	0.159	0.059	0.067	**0.161**	0.034	0.103	0.096
Biaka Pygmy	**0.134**	0.038	0.045	0.111	0.026	0.086	0.075
Bantu (Kenya)	**0.101**	0.035	0.028	0.073	0.021	0.053	0.052
Bantu (S. Africa)	**0.108**	0.049	0.027	0.066	0.023	0.063	0.055
Yoruba	**0.073**	0.030	0.021	0.050	0.023	0.034	0.040
Mandenka	**0.080**	0.025	0.027	0.065	0.022	0.042	0.046
Mozabite	**0.073**	0.027	0.019	0.045	0.023	0.035	0.039
Bedouin	**0.057**	0.038	0.020	0.040	0.019	0.028	0.035
Palestinian	**0.057**	0.033	0.017	0.040	0.020	0.023	0.033
Druze	**0.047**	0.028	0.014	0.032	0.020	0.024	0.031
Basque	0.034	**0.037**	0.009	0.022	0.021	0.018	0.027
Russian	**0.049**	0.027	0.011	0.028	0.018	0.013	0.024
Adygei	**0.038**	0.022	0.011	0.027	0.021	0.017	0.027
Balochi	**0.046**	0.036	0.015	0.038	0.020	0.021	0.031
Kalash	**0.044**	0.036	0.015	0.028	0.021	0.022	0.031
Burusho	**0.060**	0.032	0.018	0.033	0.021	0.026	0.033
Uygur	**0.048**	0.021	0.012	0.022	0.024	0.020	0.029
Yakut	**0.050**	0.034	0.015	0.027	0.025	0.019	0.032
Mongola	**0.064**	0.025	0.011	0.029	0.023	0.020	0.029
Daur	**0.057**	0.026	0.011	0.025	0.021	0.020	0.028
Yi	**0.057**	0.028	0.009	0.023	0.024	0.021	0.029
Cambodian	**0.050**	0.016	0.013	0.032	0.023	0.019	0.030
Lahu	**0.045**	0.025	0.013	0.021	0.022	0.017	0.027
Melanesian	**0.059**	0.030	0.025	0.041	0.024	0.035	0.040
Papuan	**0.068**	0.037	0.036	0.047	0.025	0.038	0.045
Pima	**0.043**	0.012	0.011	0.019	0.029	0.017	0.029
Maya	**0.032**	0.027	0.016	0.027	0.019	0.018	0.027
Colombian	**0.054**	0.051	0.013	0.031	0.023	0.020	0.032

CHAPTER IV

Haplotype Variation and Genotype Imputation in African Populations

4.1 Introduction

Africa has consistently been identified as the part of the world where the level of human genetic variation is greatest (e.g., Bowcock *et al.*, 1994; Stephens *et al.*, 2001; Ramachandran *et al.*, 2005; Tishkoff *et al.*, 2009), and genomic studies have also confirmed that African populations have the lowest levels of linkage disequilibrium (LD; Reich *et al.*, 2001; Tishkoff & Kidd, 2004; Conrad *et al.*, 2006; Jakobsson *et al.*, 2008). The high diversity and the low LD in Africa in turn influence the design and analysis of genome-wide association (GWA) studies in African populations (Rosenberg *et al.*, 2010; Teo *et al.*, 2010).

Recent strategies for finding causal variants that underlie common diseases have been based on LD, or the non-random association of variants at separate genetic loci. Because of shared inheritance of single-nucleotide polymorphism (SNP) variants at neighboring sites, an association detected between disease status and genotypes at a marker can indicate the presence of a nearby disease-susceptibility locus. Thus, highly informative "tag SNPs" that show considerable LD with other SNPs in the genome have been used as markers for finding disease associations.

The general utility of the tag-SNP approach is partly determined by the portability of tag SNPs, the extent to which tag SNPs chosen based on haplotypic patterns in a reference population perform in identifying disease genes in study populations whose patterns of haplotype variation differ from those of reference populations. Tag-SNP portability has been shown to be affected primarily by the level of LD in the study population, with genetic similarity of the reference and study populations playing a less critical but still important role (Conrad *et al.*, 2006). Thus, for populations that have relatively low levels of LD and that are genetically different from standard reference groups—a class of populations that includes much of Sub-Saharan Africa—the tag-SNP approach is less effective than for other populations.

Improved designs for GWA studies have recently used LD patterns to impute genetic variants that have not been genotyped in the study sample but that have been genotyped in a reference panel. Imputation of unknown variants, followed by testing of these variants for disease association, has been shown to improve the genomic coverage and statistical power of GWA studies (e.g., Marchini *et al.*, 2007; Servin & Stephens, 2007; Li *et al.*, 2009). Investigations of genotype imputation in worldwide populations, however, suggest that imputation accuracy is low in most African populations, again owing largely to low levels of LD and high levels of genetic diversity (Huang *et al.*, 2009a; Teo *et al.*, 2010). This difference in imputation accuracy in turn can substantially inflate the sample size required for maintaining power in imputation-based GWA studies in African populations (Huang *et al.*, 2009b).

Despite the reduced tag-SNP portability and imputation accuracy in African populations, data on patterns of haplotype variation in Africa and their applications to the design of GWA studies are relatively scarce. In this study, we extend the characterization of African haplotype diversity and LD to a total of 15 Sub-Saharan African populations, and we perform an investigation of imputation in African populations. The combination of high levels of genetic variation, low levels of LD, and large

numbers of private haplotypes in African populations makes imputation of untyped markers particularly challenging in Africans. We examine a variety of imputation designs in African populations, and by considering summary statistics on patterns of haplotype variation, we demonstrate a close relationship between maximal imputation accuracy and statistics that measure different forms of genetic similarity between samples from a target African population and those available in reference panels.

4.2 Results

4.2.1 Data

We considered a dataset of 1,107 individuals from 63 populations worldwide, including 15 Sub-Saharan African populations. Each individual was genotyped for 2,810 SNPs spread across 36 genomic regions, 16 on chromosome 21, 16 on other autosomes, and four on the non-pseudoautosomal part of the X chromosome. Each region was designed to contain a core of 60 SNPs genotyped at high density, with 12 SNPs at lower density extending in each direction away from the core. This set of genomic regions was originally chosen to represent the range of recombination rates and gene densities present in the human genome, and most SNPs were chosen among those discovered in multiethnic panels (Conrad *et al.*, 2006). The dataset subsumes the dataset of Pemberton *et al.* (2008) on 957 individuals from 55 populations (see Materials and Methods), and the 150 newly genotyped individuals (Table 4.1) represent eight Sub-Saharan African populations chosen to provide a geographically and genetically diverse subset among the samples of Tishkoff *et al.* (2009). Our investigations focus primarily on the 15 Sub-Saharan African populations.

For some analyses of imputation in study populations on the basis of external reference panels, the 1,107 individuals were augmented with genotypes in 901 unrelated individuals from 11 populations in release 2 of Phase 3 of the International

Haplotype Map Project (2010), representing a subset of the collection of 1,117 unrelated individuals in HapMap Phase 3 release 3 that was described by Pemberton *et al.* (2010). In these HapMap individuals, 517 markers were considered, all of which were located on chromosome 21 and were typed in both the 63 study populations and the 11 HapMap populations. The HapMap Phase 3 data contain four groups with significant recent African ancestry: ASW (African Americans from the southwest of the USA), LWK (Luhya from Webuye, Kenya), MKK (Maasai from Kinyawa, Kenya), and YRI (Yoruba from Ibadan, Nigeria). From the four HapMap groups, we constructed all $2^4 - 1 = 15$ possible mixtures of one or more among these four groups. We then considered each of these panels as reference data for imputation in the 15 Sub-Saharan African target populations.

4.2.2 Haplotype Variation

We assessed several aspects of haplotype variation, including "haplotype flow," private haplotypes, LD, and haplotype sharing between sampled populations and HapMap reference populations. These various computations are used later in explaining the outcomes of genotype-imputation experiments.

Haplotype flow. Using the sample-size-corrected z-statistic of Conrad *et al.* (2006), we computed pairwise haplotype sharing between major geographic regions—Sub-Saharan Africa, the Middle East (and North Africa), Europe, Central/South Asia, Oceania, and the Americas. For a fixed haplotype length, this statistic measures the fraction of haplotypes in a sample of specified size from one population that are also found in a second population. It can be viewed either as a measure of "outward haplotype flow" for the second population, quantifying the extent to which this population could have contributed haplotypes to the first population, or alternatively, as a measure of "inward haplotype flow" for the first population.

As was observed by Conrad *et al.* (2006), the outward haplotype flow from Sub-

Saharan Africa (henceforth sometimes abbreviated to "Africa") to each of the other regions exceeds the corresponding inward haplotype flow (Figure 4.1). Haplotype sharing between regions is lower when comparing Africa to other regions than when comparing most pairs of non-African regions (Figure 4.1). Consistently across haplotype lengths, haplotype sharing between Africa and other regions is greater when the full set of 15 African populations is used than when using the seven previously sampled African populations alone. It is possible that the newly sampled populations, most of which were sampled in East Africa, represent the groups that migrated out of Africa more closely than do the previously sampled groups, thereby producing increased haplotype sharing with non-Africans. Indeed, some of these populations, including Beja, Borana, and Fulani have been observed to partially cluster with Middle Eastern populations in analyses of population structure (Tishkoff *et al.*, 2009).

Private haplotypes. For each geographic region, we computed the number of private haplotypes found only in that region. Our computations used a rarefaction approach (Kalinowski, 2004; Conrad *et al.*, 2006) to adjust for differences in sample sizes across regions. We observe much larger numbers of private haplotypes in Africa than in non-African regions (Figure 4.2A), consistent with greater levels of diversity and lower LD in Africa. For example, in a sample of 54 chromosomes, for haplotypes of length 25kb, we find on average 7.35 private haplotypes in Africa, whereas we only find on average 1.71 private haplotypes in the Middle East, and even fewer in the other regions (Figure 4.2A). Within Africa, the greatest numbers of private haplotypes are found in hunter-gatherer populations, such as the San, Biaka Pygmy, and Mbuti Pygmy groups (Figure 4.2B). These three populations do not stand out in other aspects of diversity, however, as they do not have particularly large numbers of distinct haplotypes (Figure S4.1) or high haplotype heterozygosity (Figure S4.2).

Linkage disequilibrium. LD, as measured by mean r^2 values for SNP pairs in physical distance bins, declines with increasing physical distance between SNPs for

all 63 populations (Figure 4.3). African populations have the lowest levels of LD, followed by populations from the Middle East, Central/South Asia, Europe, East Asia, Oceania, and the Americas. For example, for SNPs with minor allele frequency 0.05 or greater, mean r^2 across African populations, when calculated for all SNP pairs in bins of width 6kb, drops below 0.4 at a distance of 2.5kb. The corresponding distances at which mean r^2 first drops below 0.4 are 5.2kb, 7.1kb, 9.6kb, 10.5kb, 19.2kb, and 33.3kb for the populations of the Middle East, Central/South Asia, Europe, East Asia, Oceania, and the Americas, respectively. Thus, in considering a larger sample of Sub-Saharan African populations than in most previous studies, we continue to find comparatively low LD in African populations.

Haplotype sharing with the HapMap. Using a statistic ϕ that measures the extent to which the common haplotypes in one population are also common in a second population, Conrad *et al.* (2006) found that the HapMap Phase 2 data capture common haplotypes relatively well in most groups, with the primary exception of African populations. Employing this same statistic, an expanded dataset with additional African populations, and the newer HapMap Phase 3 data, we continue to observe that for African populations, levels of sharing for common 50kb haplotypes (>10% frequency) with HapMap Phase 3 are significantly lower than corresponding levels of sharing with HapMap Phase 3 for non-African populations ($P < 0.0001$, one-sided Wilcoxon rank-sum test).

Figure 4.4 shows the fraction of common haplotypes in individual populations that are also common in the HapMap Phase 3 populations, demonstrating that the most similar HapMap group for a population is generally found in the same or the closest geographic region. Although common haplotypes of several African populations (San, Mbuti Pygmy, and Biaka Pygmy) continue to have the greatest difference from those of the individual HapMap populations, similarly to the observation of Conrad *et al.* (2006), they can generally be better captured by pooled collections consisting of two

or more HapMap Phase 3 populations than by the HapMap populations individually (Figure 4.5). In particular, testing the difference in haplotype sharing for common 50kb haplotypes in African populations with the combination panels that achieve the maximal haplotype sharing (among the 15 combinations of one or more HapMap Phase 3 populations of African descent) and with the HapMap Phase 3 YRI panel, sharing is significantly greater with the combination panels than with the YRI panel alone ($P < 0.0001$, one-sided Wilcoxon signed-rank test).

4.2.3 Genotype Imputation

To understand the properties of genotype imputation in African populations, we considered two designs, both using the software MACH (Li *et al.*, 2006, 2010). We first examined imputation accuracy for all pairs among the 63 populations, with one population chosen as the reference and another as the target. We next identified, for each of the 15 African populations, the optimal reference panel chosen from the HapMap.

Imputation at untyped markers based on population samples. To examine the variation in imputation accuracy across potential reference populations, for each of 63×63 population pairs consisting of a target population and a reference population, we imputed missing genotypes at randomly selected hidden markers in the target population on the basis of a small panel of individuals in the reference population, holding reference panel size constant at six individuals. The panel size of six individuals corresponds to the smallest sample size among all 63 populations, and therefore, it represents the largest panel size that permits comparable evaluations of all pairs of distinct populations.

Considering all 63×63 imputations, we find that except for African target populations, imputing missing genotypes in a target population on the basis of a reference population from the same geographic region yields higher imputation accuracy than

the mean of all values in the 63×63 matrix of imputation accuracies (Figure 4.6). By contrast, imputing missing genotypes in African target groups using non-African reference groups yields imputation accuracy lower than the mean, except in a few target populations (e.g., Beja, Iraqw, and Sandawe with the Mozabite group as reference). Among all 779 pairs consisting of reference and target populations from the same geographic region, we find that 30.4% of the imputations appear in the top 10% of all 63×63 imputation accuracies, with values ranging from 88.2% to 94.6%. On the other hand, among 720 pairs consisting of an African target population and a non-African reference population, 36.7% appear in the bottom 10% of imputation accuracies, with values ranging between 59.3% and 78.2%.

In this imputation experiment, we observe an asymmetry of imputation performance in population pairs consisting of a reference population and a target population with different geographic origins. That is, in many cases, imputation using one population as a reference panel and a second population from a different geographic region as a target has considerably higher or lower accuracy than in a scenario with the roles of the populations reversed. This reference-target asymmetry is most pronounced in population pairs in which one population is African and the other is non-African; in 628 or 87.2% of 720 such pairs (15 African $\times$ 48 non-African populations), imputation accuracy is lower when imputing untyped markers in an African population on the basis of a non-African population than when performing imputation in the reverse direction. For population pairs of non-African descent from different geographic regions, we observe a similar reference-target asymmetry. For instance, in 113 or 78.5% of 144 pairs containing a European and an East Asian population (8 European $\times$ 18 East Asian populations), imputation accuracy is lower in the European population than in the East Asian population on the basis of the other population as reference data.

Evaluating the portability of a reference population for imputation in target popu-

lations other than the reference population itself, we consider two metrics—the number of target populations in which a reference population serves as either the best or second-best reference panel, and the mean imputation accuracy across target populations in which imputation is performed using the reference population. Using the first metric to identify top-performing reference groups across the range of possible target populations, we find Sengwer and Yoruba to be the most portable reference groups for imputation in African populations. Sengwer is the best or second-best reference group in six of the 14 other African samples, and Yoruba is the best or second-best panel in five of 14. Additionally, Sengwer and Yoruba produce the highest mean imputation accuracy across the 14 remaining African populations (86.0% and 85.8%, respectively).

Imputation at untyped markers based on the HapMap. To identify suitable HapMap reference panels for imputation in the 15 African populations, in each population, we masked a fixed set of randomly selected markers and then imputed missing genotypes at these markers on the basis of each of the 15 possible combinations of the four HapMap panels of African descent.

For each African target population, Figure 4.7 reports the optimal reference panel chosen from the 15 combinations of HapMap reference groups. All except one of the African populations are most accurately imputed using a reference panel that contains individuals from new HapMap Phase 3 samples of African ancestry (ASW, LWK, and MKK). The only exception is Mandenka, for which the optimal reference panel consists solely of the HapMap YRI population. The combined panel of all four HapMap populations of African origin is not the optimal reference group in any of the 15 African populations, and it is the second-best reference panel in only three of the 15 African groups (Kenyan Bantus, Fulani and Mada). Interestingly, several populations (Beja, Biaka Pygmy, Borana, Fulani, Mbuti Pygmy, and Sandawe) have in their optimal reference panels the HapMap ASW admixed sample of African Americans.

On the basis of reference panels consisting of mixtures of the HapMap Phase 3 populations, the San, Mbuti Pygmy, and Biaka Pygmy populations continue to be the most poorly imputed groups, as was previously observed with earlier reference panels from HapMap Phase 2 (Huang *et al.*, 2009a). Yoruba remains the best-imputed population, with the combination of the HapMap LWK and YRI populations as its optimal reference panel. Although the size of the underlying optimal reference panels varies widely across the 15 target populations, from 80 individuals for the LWK panel to 284 individuals for the combined panel containing the HapMap LWK, MKK, and YRI populations, maximal imputation accuracy varies only moderately across the 15 African target populations. The highest and lowest values differ by less than 7.0% among all 15 populations, and by less than 2.0% for the 11 populations with highest maximal imputation accuracy.

To evaluate the improvement in imputation accuracy in African populations resulting from the addition of the ASW, LWK, and MKK samples to the HapMap Phase 3 data, for each African population, we computed the difference between the maximal imputation accuracy in the population using its optimal combination of reference panels and the imputation accuracy in the population on the basis of the YRI reference panel. Averaged across African populations, the increase in imputation accuracy is 1.3%, corresponding to a mean percentage reduction of 11.1% in imputation error rates. Note, additionally, that the HapMap Phase 3 YRI panel examined in our study contains 80% more unrelated individuals compared to the HapMap Phase 2 YRI panel (from 60 to 108 unrelated individuals); this panel is thus likely to produce higher imputation accuracy than the earlier panel. Consequently, as a measure of the improvement in African imputation accuracy on the basis of HapMap Phase 3 compared to HapMap Phase 2, our estimate is likely to be conservative.

To further quantify contributions of individual HapMap Phase 3 panels of African ancestry to imputation accuracy in the 15 African populations, for each HapMap

panel of African origin, we computed the difference in maximal imputation accuracy attainable in each of the 15 populations using two optimal reference panels, one chosen from a full collection of combination panels and the other chosen from a reduced collection. The full collection consisted of all $2^4 - 1 = 15$ combinations of the four HapMap Phase 3 panels, producing the maximal imputation accuracies shown in Figure 4.7. The reduced collection, a subset of the full collection, consisted of $2^3 - 1 = 7$ combinations of the same panels of African descent, excluding the panel whose contributions were under evaluation. A larger difference in maximal imputation accuracy, examining the full and reduced collections, suggests a greater impact of the HapMap panel under consideration, because of a greater difference in imputation accuracy achieved with and without the panel. For each of the 15 African populations, we ranked the four HapMap Phase 3 panels of African origin by the difference in maximal imputation accuracy, finding that the HapMap ASW panel has the greatest influence on maximal imputation accuracy only in Fulani, a group that has been suggested to have had recent gene flow both with Sub-Saharan African and with Eurasian populations (Scheinfeldt *et al.*, 2010). Considering the remaining 14 African populations, exclusions of the HapMap MKK, LWK, and YRI panels produce the greatest impact in six, five and three populations, respectively. Among the target populations whose imputation accuracies are most strongly influenced by a particular panel, the mean percentage reductions in imputation error rates are 4.1%, 10.4% and 8.3% for MKK, LWK and YRI, respectively (the percentage reduction in imputation error in Fulani when including the ASW reference panel is 3.8%).

Relating imputation to haplotype variation. The selection of optimal reference panels for imputation in target populations generally requires an investigator either to have prior knowledge of the performance of candidate panels in the target populations or to perform imputation experiments similar to the ones described in the preceding two sections. However, prior knowledge might be unavailable for unusual

target populations, and imputation experiments can be computationally intensive. Thus, for target populations that have not been the focus of previous imputation studies, the ability to predict the optimal reference panel among a collection of candidate panels on the basis of simple genotypic and haplotypic variation statistics computed for the target and each of its candidate reference groups, can serve as a computationally attractive approach to the selection of reference panels.

To provide a basis for predicting properties of imputation from statistics on variation patterns, we examined the dependence of imputation-accuracy results (Figure 4.7) on our analysis of haplotype variation in the 15 African populations. Both imputation accuracy and haplotype variation were investigated using the same set of 517 markers that overlapped between our study populations and the HapMap Phase 3 populations. We considered three haplotype-variation statistics from the Haplotype Variation section (haplotype sharing for a target population with a reference population, number of private haplotypes in the target population, and level of LD in the target population), as well as F_{st} between target and reference populations, as possible predictors of imputation accuracy in a target population on the basis of a reference population. Haplotype sharing and F_{st} are reasonable predictors because they provide measures of genetic similarity and distance between a target group and a reference group. The number of private haplotypes provides a measure of the distinctiveness of a target population and thus might be expected to be inversely related to imputation accuracy. Lastly, the level of LD as measured by r^2 is a reasonable predictor because the strength of correlation among nearby SNPs on a target haplotype underlies our ability to impute genotypes at an untyped SNP using genotype information at a nearby typed SNP.

For the 15 African target populations, with missing genotypes imputed based on their respective optimal HapMap mixtures, Figure 4.8 displays the relationships of imputation accuracy with four summary statistics—the number of private haplotypes

of length 50kb, the level of LD at 50kb, the haplotype sharing for target populations with their optimal reference groups using a window size of 50kb, and F_{st} between the target and reference populations. Haplotype sharing for a target population with a reference population, as well as F_{st} between a target population and a reference population, each produce a strong relationship with imputation accuracy in the target (with Pearson correlation coefficient $r = 0.79$ and $P = 0.0004$ between imputation accuracy and haplotype sharing, and $r = -0.86$ and $P < 0.0001$ between imputation accuracy and F_{st}). The relationship between imputation accuracy and the number of private haplotypes is weaker ($r = -0.66, P = 0.0070$), and the relationship between imputation accuracy and the level of LD is not statistically significant ($r = 0.15, P = 0.6044$).

Statistics on genetic similarity between an African target population and a HapMap reference group can in some cases be used for identifying the optimal reference panel for imputation in the target. Each plot in Figure 4.9 shows the imputation accuracies in a given target population on the basis of each of the 15 HapMap mixture panels, sorted on the x-axis according to the haplotype-sharing statistic. In four of 15 target populations, the optimal HapMap mixture, as shown in Figure 4.7, is indeed the mixture with the highest haplotype sharing; in most target populations, use of the mixture with the highest haplotype sharing leads to a relatively small decrease in imputation accuracy compared to use of the optimal mixture. For each target population, we computed the difference in accuracy between the imputation performed using the mixture with the highest value of the haplotype-sharing statistic and the imputation performed using the optimal HapMap mixture. The mean loss of imputation accuracy across the 15 African target populations in this case is 0.0038, corresponding to a mean percentage increase of 4.2% in imputation error.

Similarly, each plot in Figure 4.10 shows the imputation accuracies in a target population on the basis of the 15 HapMap mixture panels, sorted instead on the x-axis

according to F_{st}. The optimal HapMap mixture is the mixture with the lowest F_{st} in only 3 of 15 target populations. However, in many of the remaining target populations, the imputation accuracy obtained using the mixture with the lowest F_{st} is only very slightly lower than the imputation accuracy obtained using the optimal mixture. The mean loss in imputation accuracy from use of the lowest-F_{st} mixture rather than the optimal mixture is 0.0013, corresponding to a mean percentage increase of 1.3% in imputation error. This small difference in imputation accuracy suggests that genetic similarity between target and reference populations plays a central role in predicting imputation accuracy in the target population, and that similarity statistics can be used to guide the selection of suitable reference populations.

4.3 Discussion

Genotype imputation has played an increasingly important role in the analysis of human genetic variation and genotype-phenotype association, and the continuing growth of genomic resources facilitates the expansion of imputation studies into new populations. We have found that the availability of additional HapMap Phase 3 populations of African descent increases the accuracy of genotype imputation in Sub-Saharan African populations, improving the prospects for GWA studies in these groups. Focusing on populations from Sub-Saharan Africa, we have presented a detailed investigation of haplotype diversity and genotype imputation, recommending the use of haplotype-sharing measures and F_{st} between a target population and candidate reference populations as guiding criteria for selecting reference panels for imputation in the target population.

We characterized the level of genetic similarity between populations by the magnitude of their haplotype sharing. Examining the patterns of haplotype sharing at a regional level, we confirmed earlier observations of asymmetry between African and non-African populations in haplotype sharing, as reflected in the greater "outward"

than "inward" haplotype flow from Africa to other geographic regions (Conrad *et al.*, 2006). This asymmetry in haplotype sharing (Figure 4.1) provides a partial explanation for a corresponding reference-target asymmetry in imputation performance for Africans and non-Africans (Figure 4.6). In particular, the net outward haplotype flow from Africa to other geographic regions implies that for a non-African haplotype targeted for imputation on the basis of an African reference population, the probability of finding the same haplotype inherited by descent in the reference population is greater than the probability of finding an African haplotype targeted for imputation in a non-African reference population. An increased probability of finding reference chromosomal stretches inherited by descent for a non-African target haplotype in turn produces an increased probability of correctly inferring missing genotypes of the non-African target on the basis of African reference haplotypes, compared to the probability of correctly inferring missing genotypes of an African target on the basis of non-African reference haplotypes. Following the same argument, we can attribute much of the asymmetry in imputation performance between collections of populations from different geographic regions to the asymmetry in haplotype sharing for the populations involved.

The accuracy with which genotypes can be imputed in a target population, though positively correlated with haplotype sharing and the F_{st} statistic with the reference panel, is clearly not solely determined by either of these measures of genetic similarity between target and reference populations. For example, considering the 15 African populations, the Mandenka population had the highest maximal haplotype-sharing fraction across the 15 possible mixtures of the HapMap Phase 3 populations of African descent (Figure 4.5). Among the 15 African target populations, however, the Mandenka population had less than the median maximal imputation accuracy on the basis of the optimal reference panel chosen among the 15 HapMap mixtures. Future theoretical work will be important for clarifying the determinants of imputation

accuracy; in the absence of such work, further investigation of empirical approaches, some inspired by population-genetic theory, can continue to provide improvements to imputation in novel target populations (e.g., Egyud *et al.*, 2009; Huang *et al.*, 2009a; Li *et al.*, 2010; Paşaniuc *et al.*, 2010; Shriner *et al.*, 2010).

Although our dataset in 63 worldwide populations enables us to investigate factors affecting accuracy of genotype imputation in diverse populations, especially in Sub-Saharan Africans, the relatively small numbers of markers and sample sizes do limit the scope of our study. For example, because of the small size of the marker set, the fraction of the markers that we chose to impute in our experiments was less than that typically used in GWA applications, for which larger fractions of the dataset are imputed rather than genotyped. This small size of the marker set had the additional consequence that in our imputation experiment involving the HapMap, for each of the 15 African target populations, imputation accuracies resulting from use of the top choices of reference panels did not differ substantially, thereby limiting our ability to provide clear support for particular mixtures of HapMap panels (Figure 4.7). Further, for our 63×63 imputation experiment involving only data from the 63 populations, we relied on phased haplotypes, and relatively small sample sizes might have limited phasing accuracy; because phasing accuracy is lowest in populations with lower LD (Conrad *et al.*, 2006), phasing errors could have contributed to the elevated imputation error rates in African target populations (Figure 4.6). We also note that while the MACH software that we used is among the most commonly used imputation programs, other methods such as BEAGLE (Browning & Browning, 2007, 2009) and IMPUTE (Marchini *et al.*, 2007; Howie *et al.*, 2009) are frequently employed. While the numerical results of the imputation experiments would likely vary with our methodological choices, however, our primary goal has been to examine the way in which imputation accuracies relate to each other across different reference and target populations, with a focus on Sub-Saharan Africa, and we do not expect that these

general patterns would be substantially affected by changes to the imputation software, marker sets, or sample sizes. The limitations of our imputation experiments will become easier to address as large-scale African population-genetic datasets proliferate, from such sources as genomic studies of human evolution (e.g., Bryc *et al.*, 2010; Henn *et al.*, 2011) and GWA studies in African and African-American populations (e.g., Adeyemo *et al.*, 2009; Jallow *et al.*, 2009; Teo *et al.*, 2010).

4.4 Materials and Methods

4.4.1 Data

SNP data. We supplemented the worldwide set of 957 individuals studied by Pemberton *et al.* (2008), which itself updated the dataset of Conrad *et al.* (2006) on 927 individuals from 53 populations, with data on eight additional African populations. Among 160 African individuals genotyped initially, four were discarded as a result of poor genotyping quality. For each pair among the remaining 156 individuals, the fractions of SNPs at which the pair shared 0, 1, and 2 identical alleles were calculated. The computation used all SNPs at which genotyping was attempted, and it identified two pairs of duplicate samples and five pairs of close relatives, two of which shared one individual. This shared individual was removed from both pairs, and from each of the five remaining pairs, the individual with the greater amount of missing data was removed. Research and ethics approvals and permits were secured prior to sample collection, as detailed by Tishkoff *et al.* (2009). Written informed consent was obtained on-site from all participants, and the institutional review boards of the University of Maryland at College Park and the University of Pennsylvania approved the study.

Genotyping was attempted for the African individuals at 3,024 SNPs spread across 36 genomic regions, simultaneously with genotyping of the 30 Indian samples that

formed the focus of the work of Pemberton *et al.* (2008). The preparation of the final dataset for this study appears in Pemberton *et al.* (2008), who incorporated the African samples in producing a final dataset of 2,810 SNPs, but then omitted these samples in data analysis. Our final dataset, considering all 1,107 individuals and 2,810 SNPs, has a missing data rate of 0.11% (0.38% in the 150 newly sampled African individuals). Of the 2,810 SNPs, a subset of 1,272 SNPs are located on chromosome 21. To investigate genotype imputation in our study samples, we focused on this subset, which has a missing data rate of 0.10% (0.36% in the 150 newly sampled African individuals). Haplotype phasing utilized fastPHASE 1.0 (Scheet & Stephens, 2006), following the same approach as in Conrad *et al.* (2006), and it was completed by Pemberton *et al.* (2008).

HapMap data. For some analyses, we incorporated additional reference individuals for genotype imputation. The reference data consisted of 901 unrelated individuals in release 2 of HapMap Phase 3 (2010). We used a dataset in which phased genotypes in these individuals were available at 1,361,534 autosomal SNPs. Of these SNPs, 18,943 were on chromosome 21, among which 517 were also available in the 1,107 study individuals. For imputation designs involving HapMap individuals as reference data, we assessed imputation accuracy at a subset of the 517 SNPs by using the unphased genotypes at the 1,272 SNPs from the study sample and the phased genotypes at the 18,943 chromosome-21 SNPs in the HapMap Phase 3 data. For imputations that instead used populations in the study sample as reference data, we evaluated imputation accuracy at a subset of the 1,272 SNPs, using unphased data for target samples and phased data for reference samples at those SNPs.

4.4.2 Statistical Analyses of Haplotype Variation

Haplotype windows. We computed haplotype summary statistics using haplotypes defined by "core" SNPs in genomic windows of size w base pairs. In the set

of SNPs genotyped, core SNPs are SNPs that lie within a more densely genotyped region with a mean spacing of ∼1.5kb between consecutive SNPs (non-core SNPs lie in the flanking regions of each core, with a mean spacing of ∼10kb). For each SNP in a "core" region, a haplotype locus is specified by the set of allelic states at all SNPs located in the half-open window $[a, a + w)$, where a denotes the position of the SNP under consideration and $a + w$ denotes the position along the chromosome w base pairs away from the position a. All SNPs defining a haplotype locus are required to lie completely within a core region. Furthermore, identical haplotypes must have the same variants for all SNPs with positions in $[a, a + w)$. For each value of the window size w, we present summary statistics averaged across all haplotype loci of size w. For instance, for a given population, haplotype heterozygosity was computed for each haplotype locus and was then averaged across haplotype loci.

Unless otherwise noted, summary statistics on haplotype variation were calculated twice in our study. We first computed the statistics using all 1,800 core SNPs outside X-chromosomal regions (numbered 23-26 in Table SM.2 of Conrad *et al.*, 2006) for the characterization of haplotype variation in the study populations (Figs. 4.1-4.3). The collection of 1,800 SNPs was identical to that used by Pemberton *et al.* (2008). For the investigation of the relationship between haplotype variation and imputation performance (Figs. 4.4, 4.5, and 4.8), we repeated the computation using the set of 517 SNPs that overlapped between the study samples and the HapMap Phase 3 data so that results on haplotype variation and on imputation accuracy used the same underlying set of SNPs. Lastly, we computed pairwise F_{st} between each of 15 African target populations and each of 15 mixtures of HapMap Phase 3 panels of African ancestry, using the set of 517 SNPs and eq. 5.3 of Weir (1996). All haplotype summary statistics, as well as F_{st}, were computed using phased datasets.

Numbers of distinct haplotypes and private haplotypes. To adjust for sample-size differences across populations and geographic regions, following Conrad

et al. (2006), we used a rarefaction approach for estimating the numbers of distinct haplotypes and private haplotypes. For each of these two statistics, in a sample of size N, this approach chooses a value $g \leq N$ and it obtains the statistic by averaging the expected value of the statistic across all possible subsamples of size g from the original sample of size N. This method enables a correction for differing sample sizes across populations, as the same value of the subsample size g can be used in evaluating a statistic in each population. For all population-level computations of the two statistics, we used $g = 12$, which corresponds to the smallest sample size among 63 populations. For all computations involving geographic regions, we used $g = 54$, as the smallest sample size among the seven geographic regions equaled 54 chromosomes.

Haplotype sharing. To compute the fraction of distinct haplotypes shared between two populations, j and j', we used the z-statistic of Conrad *et al.* (2006). For each haplotype locus, we first computed the numbers of distinct haplotypes and the numbers of private haplotypes for each of the two populations, where private haplotypes for population j refer to those not found in population j'. This computation used rarefaction with $g = 54$ when comparing geographic regions and $g = 12$ when comparing populations.

The expected number of distinct haplotypes found in a sample of size g from population j that will also be found in a sample of size g from population j' is then equal to the difference between the expected number of distinct haplotypes in population j and the expected number of private haplotypes in population j. Thus, the z-statistic of Conrad *et al.* (2006) is an estimator of the fraction of distinct haplotypes observed in a sample of size g from population j that will also be observed in a sample of size g from population j'.

Linkage disequilibrium. We measured LD by the correlation coefficient, r^2, between all pairs of SNPs with minor allele frequency greater than some cutoff value,

c, where $c \in [0, 1)$. For each population, we computed the mean r^2 and the mean distance between pairs of SNPs for all SNP pairs within bins of size b; a bin centered on distance x contains all pairs of distinct SNPs in the interval $(x - b/2, x + b/2]$. We tested the sensitivity of r^2 values to various choices of c (0, 0.05, and 0.1) and b (1 kb, 3 kb, 6 kb, and 10 kb), and we found that the choices of c and b had relatively little effect on the observed LD patterns.

Haplotype sharing with the HapMap. Using the ϕ statistic (Conrad *et al.*, 2006), for each population, we computed the fraction of haplotypes common in a population that were also common in each of the 11 HapMap Phase 3 populations and in the 15 combinations of one or more HapMap Phase 3 groups of African descent. This statistic evaluates the number of distinct haplotypes that are common in each of a pair of populations, as a fraction of the number of distinct haplotypes common in the population from the pair designated as the "donor." We used $g = 12$ in rarefaction-based evaluations of the number of distinct haplotypes, and the set of 517 SNPs that overlapped with the HapMap Phase 3 data was used for computations of ϕ. Estimates of ϕ were generally insensitive to the choice of cutoff used for defining "common" haplotypes (haplotype frequency >0.01, >0.05, or >0.1). The ϕ statistic was obtained by averaging across haplotype loci within each of the genomic core regions, and it was then averaged across genomic regions.

4.4.3 Genotype-imputation Experiments

Imputation at untyped markers based on population samples. We examined how well missing genotypes in each population can be imputed using other population samples as reference panels. For each population in which imputation was performed, we masked the same set of 77 SNPs on chromosome 21, randomly chosen among the 517 markers that overlapped between our samples and the HapMap Phase 3 populations. We then estimated genotypes at these markers using the soft-

ware MACH (Li *et al.*, 2006, 2010). MACH settings were identical to those used in imputations of untyped markers in Huang *et al.* (2009a) except that we dropped two options, `interimInterval`, which outputs intermediate results, and `mask`, which masks a specified proportion of genotypes (as opposed to masking the genotypes of specific markers in all individuals). For improved genotype estimates, we also increased `rounds`, the number of rounds for the Markov sampler, from 20 to 50. The median minor allele frequency of the 77 hidden SNPs ranges from 0.1957 to 0.2895 across the 15 African populations, and from 0.1875 to 0.3036 across 61 populations (the median minor allele frequency is lower in the Surui and Pima populations).

In each target population, imputation was performed 62 times, each time based on a subset of the unmasked, phased data from one of the remaining populations as a reference group. The target data of a population consisted of unphased genotype data in all individuals available from that population. For all target populations, we used the same reference data, consisting of haplotypes of six individuals randomly selected from a reference population.

Additionally, we imputed each population on the basis of itself. For each population, we split its data into two non-overlapping sets and used one set to impute the other. For 61 of 63 groups, we used the same reference sets of six individuals described above. For two population samples of size six individuals (San and Tuscan), we randomly selected five instead of six individuals and created the reference set using the unmasked, phased genotype data of these individuals. We then used unphased genotype data for individuals not sampled for inclusion in the reference set to form the target set for the evaluation of imputation accuracy. Thus, for imputation in a target population with sample size n using reference data from the same population, for 61 populations, the target set consisted of $n - 6$ individuals that were not in the reference set and for the remaining two populations, it contained the unique individual that was not in the reference set.

Lastly, to summarize imputation performance in each population, we estimated allelic imputation accuracy using eq. 1 of Huang *et al.* (2009b), which employs MACH-estimated genotype posterior probabilities and averages them across SNPs and across individuals in the target population sample. Imputation error is then defined as one minus imputation accuracy. We averaged imputation accuracy across 10 replicates of our imputation experiment, each time using one of ten randomly selected sets of reference individuals (the mean across the replicates is plotted in Figure 4.6).

We note that except in three African populations (Iraqw, Sengwer, and Borana) that have slightly elevated native missing data rates of 0.42%, 0.67% and 0.71%, the other 60 populations have similarly low rates of natively missing data, ranging between 0.01% and 0.29% across the 1,272 markers on chromosome 21 ("natively missing data" refer to data missing prior to our intentional masking of SNPs in the experimental design; all natively missing data rates were computed using unphased subsets of our final dataset).

Imputation at untyped markers based on Hapmap populations. We next evaluated the use of HapMap Phase 3 populations as reference data and identified optimal reference panels for imputing missing genotypes in the various African populations. The same collection of 77 SNPs (∼15% of 517 overlapping SNPs between the HapMap data and our data) masked in the previous experiment was masked, and the unphased genotypes of these hidden SNPs were estimated using identical MACH settings to those in the previous section, except that we modified the "`seed`" option to change the initial random seed used by MACH from its default value of 123456. The values plotted in Figure 4.7 were obtained as means across 10 replicates, with the replicates having varying random seeds for the MACH runs. We considered as reference data combinations of HapMap Phase 3 groups of African origin, pooling phased genotypes of unrelated individuals from the four populations with significant recent African ancestry (40 ASW, 80 LWK, 96 MKK, and 108 YRI individuals). In

total, $2^4 - 1 = 15$ possible combinations were considered. Because we combined the panels with their original sizes, the 15 combination panels varied in size. Imputation accuracy was assessed in the same manner as in the previous experiment.

Relating imputation to haplotype variation. To explore the relationship between imputation accuracy and summary statistics on genotypic and haplotypic variation, we investigated the correlation between maximal imputation accuracy in the 15 African populations on the basis of the optimal panel chosen among the 15 HapMap combinations and each of several summary statistics: number of private haplotypes, LD as measured by r^2, haplotype sharing as measured by the fraction of common haplotypes also found in the optimal panel among the 15 choices, and F_{st} between a target population and its corresponding optimal mixture of the HapMap Phase 3 panels. The number of private haplotypes and the fraction of common haplotypes shared with the HapMap were computed using a window size of 50kb. Values of r^2 were determined using 6kb bins, and F_{st} was computed for individual SNPs and then averaged across SNPs. Imputation and haplotype-variation results were obtained using the same underlying set of 517 SNPs that overlapped between the HapMap data and our study samples. We computed the Pearson correlation coefficients between imputation accuracy and each of the four statistics.

4.5 Web Resources

HapMap Phase 3 data,
`http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009-02_phaseIII/HapMap3_r2/`
MACH software, `http://www.sph.umich.edu/csg/abecasis/MACH/`

Population	Sampling location	Language family	Sample size
Beja	Sudan	Afroasiatic	20
Borana	Kenya	Afroasiatic	18
Fulani	Cameroon	Niger-Kordofanian	19
Hadza	Tanzania	Khoesan	18
Iraqw	Tanzania	Afroasiatic	18
Mada	Cameroon	Afroasiatic	19
Sandawe	Tanzania	Khoesan	20
Sengwer	Kenya	Nilo-Saharan	18

Table 4.1: Eight newly genotyped African populations incorporated in the study. The Beja and Fulani samples are from the Tishkoff *et al.* (2009) Hadandawa Beja and Mbororo Fulani samples, respectively.

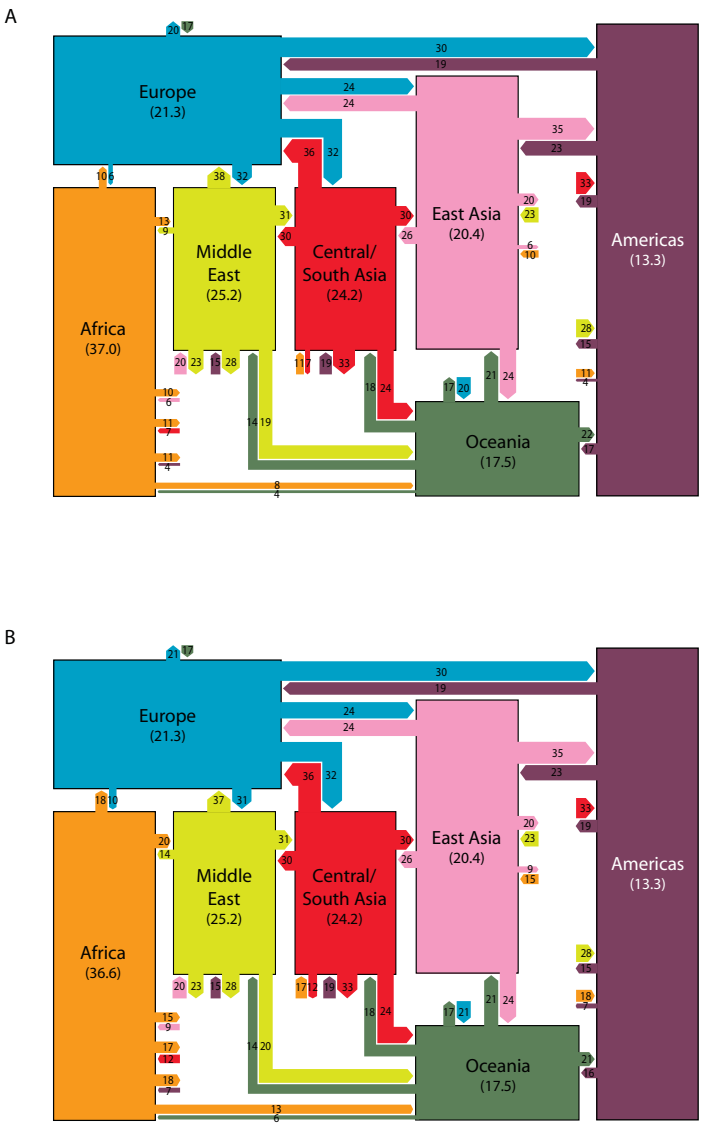

Figure 4.1: Schematic world map of haplotype variation. (A) Haplotype sharing on the basis of the data from Pemberton *et al.* (2008). (B) Haplotype sharing after including eight newly sampled African populations. The mean number of haplotypes per genomic core region in a sample size of 54 chromosomes is written for each geographic region. Links entering a geographic region indicate the percentages of distinct haplotypes from the geographic region found in other regions and are drawn proportionately in width. For example, in part A, on average 10% of haplotypes observed in Europe are found in Africa (18% in part B), whereas 6% of African haplotypes are found in Europe (10% in part B). The links can be viewed as a description of haplotype "flow": for example, 10% (18%) gives a measurement of the proportion of distinct European haplotypes that could have come from Africa (without mutation or recombination), and 6% (10%) gives the proportion of African haplotypes that could have come from Europe. We used 1,800 core SNPs to generate the figure.

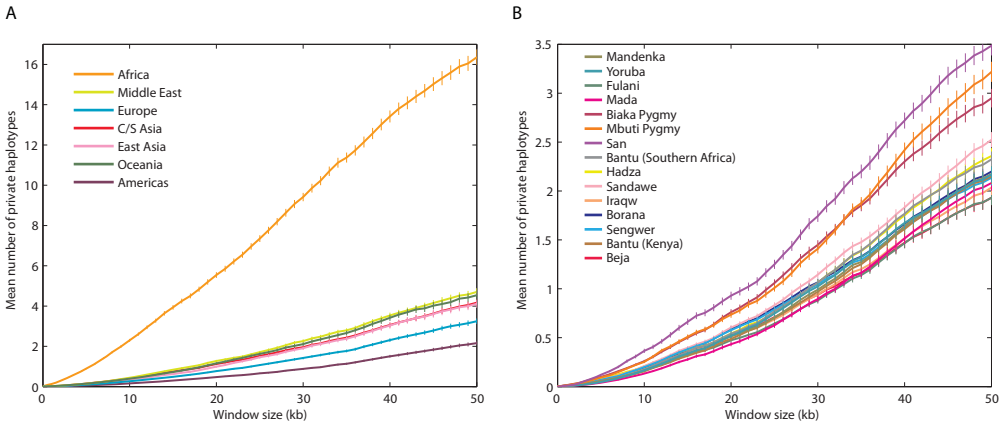

Figure 4.2: Numbers of private haplotypes. (A) The number of private haplotypes in each geographic region as a function of haplotype length. Sample sizes were adjusted to represent 54 chromosomes from each geographic region. (B) The number of private haplotypes in each African population as a function of haplotype length. Sample sizes were adjusted to represent 12 chromosomes from each population. Error bars represent the standard error of the mean across haplotype-loci.

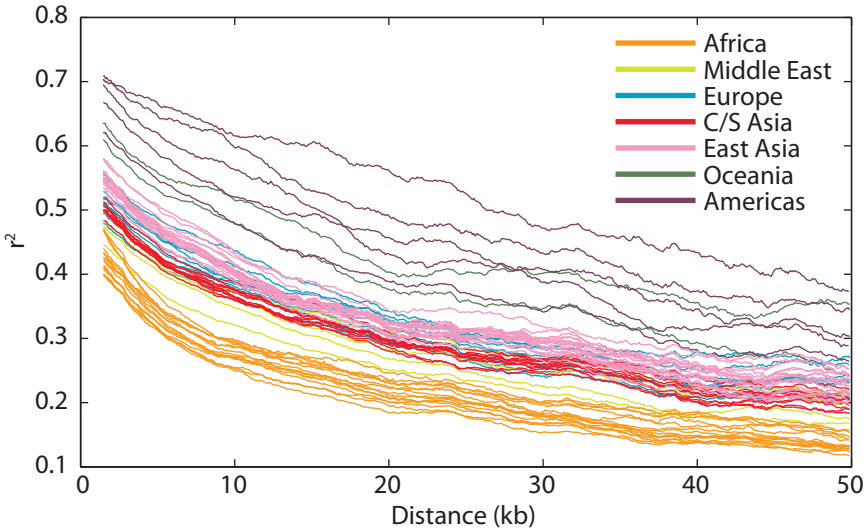

Figure 4.3: Linkage disequilibrium vs. physical distance. r^2 was calculated for each pair of SNPs with minor allele frequency greater than or equal to 0.05. The mean r^2 within a bin is plotted as a function of the mean of the distance between pairs of SNPs within the bin. The bin size was 6kb. Lines for individual populations are color-coded by geographic region.

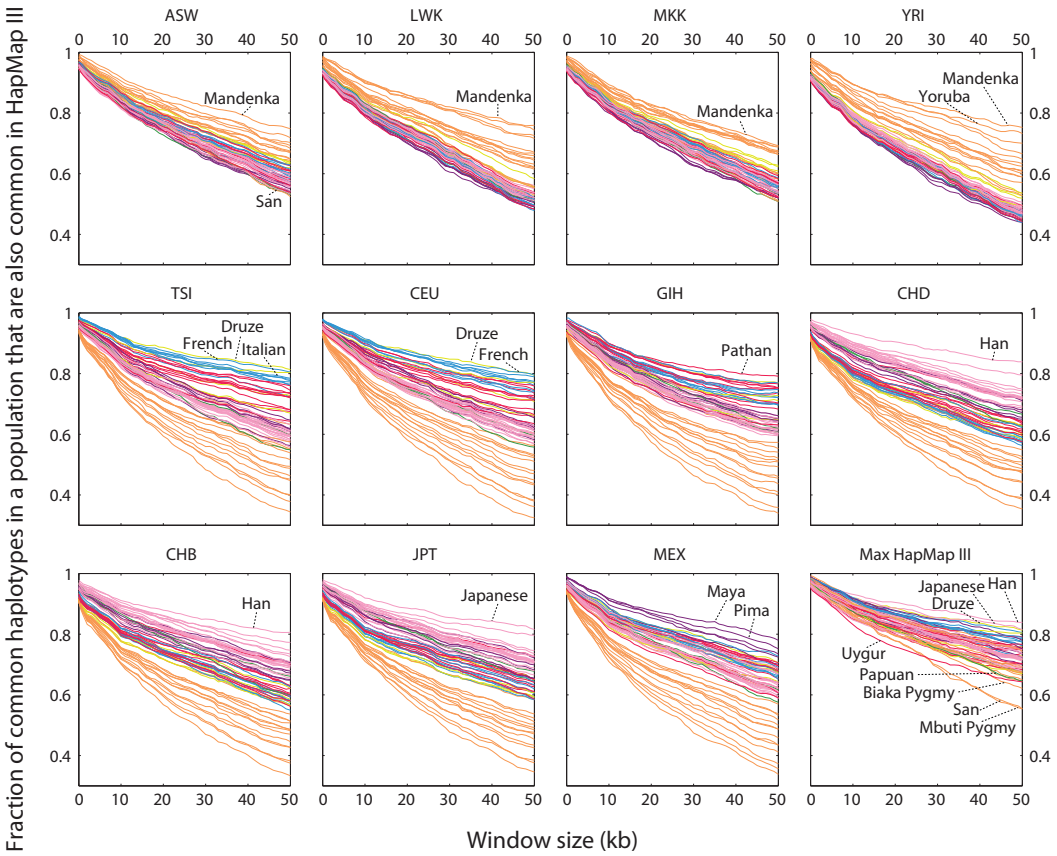

Figure 4.4: The fraction of common haplotypes in individual populations that are also common in the HapMap. For each plot we used haplotypes based on the 517 SNPs that overlap between HapMap Phase 3 and our autosomal core regions on chromosome 21. We first averaged over all haplotype-loci within each core region and then averaged across the core regions for windows of a given length. Each curve shows the fraction of the common haplotypes of a population (with >10% frequency) that are also common in a HapMap sample.

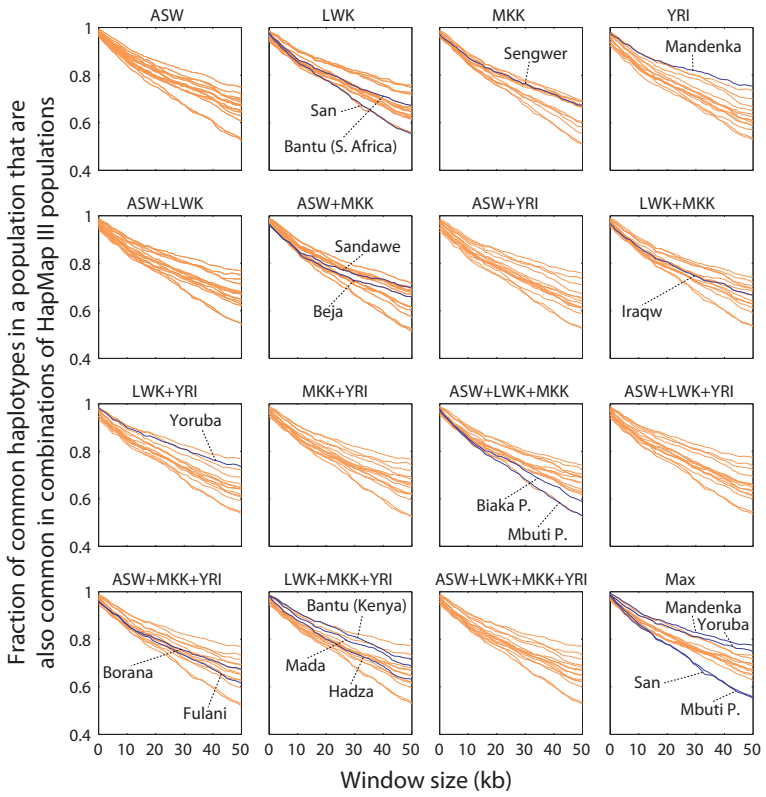

Figure 4.5: The fraction of common haplotypes in African populations that are also common in the HapMap. For each plot we used haplotypes based on the 517 SNPs that overlap between HapMap Phase 3 and our autosomal core regions on chromosome 21. We first averaged over all haplotype-loci within each core region and then averaged across the core regions for windows of a given length. Each curve shows the fraction of the common haplotypes of a population (with >10% frequency) that are also common in a HapMap sample formed by combining specific HapMap groups with recent African ancestry. Inside each plot that corresponds to one of the 15 HapMap mixtures, we label target populations in which the corresponding HapMap mixture served as the optimal reference panel among the 15 mixture panels. For the last plot of maximal haplotype sharing with the HapMap, we label the populations with the highest and lowest maximal sharing fractions.

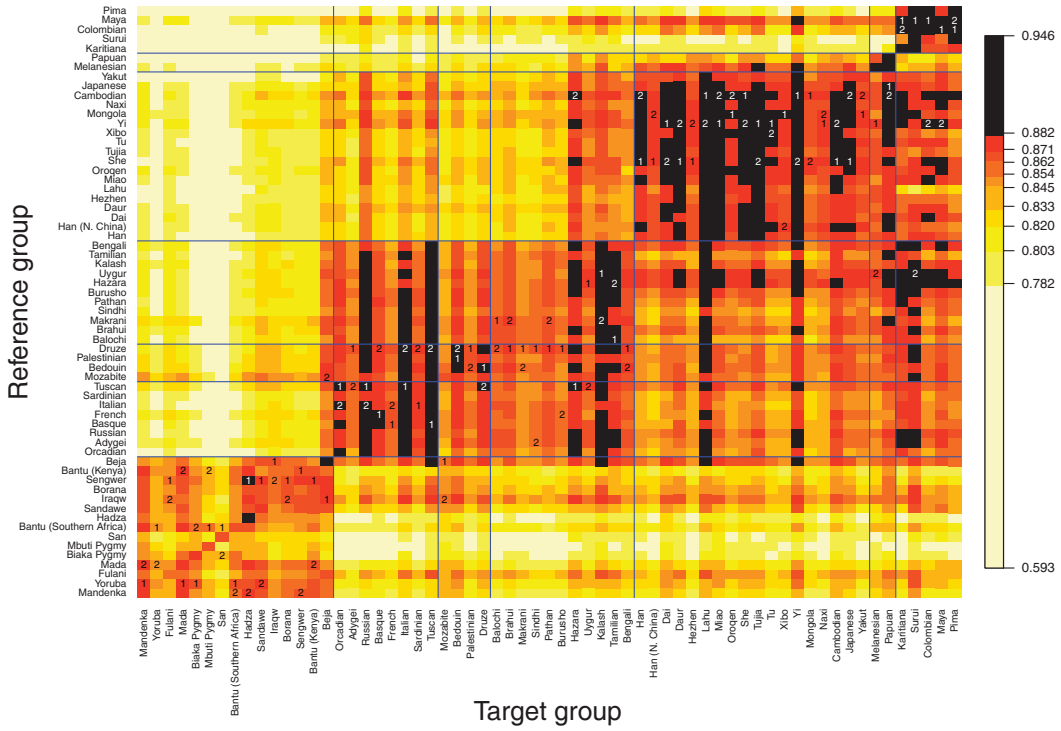

Figure 4.6: Imputation accuracy for inference of genotypes at hidden markers. For each target population specified by the column label, we masked a set of markers and imputed genotypes in the population using the reference population specified by the row label. Of 1,272 markers, 77, or ∼6%, were randomly chosen among a subset of 517 markers and masked, and for each target, the same set was masked for imputation with each reference population. The colors correspond to ten deciles of imputation accuracy across all populations and all reference panels. For each population, the best and second-best reference panels among 62 other populations are labeled 1 and 2, respectively. For convenience in interpreting the figure, the horizontal and vertical blue lines separate results by geographic region (from left to right and from bottom to top: Africa, Europe, Middle East, Central/South Asia, East Asia, Oceania, and the Americas).

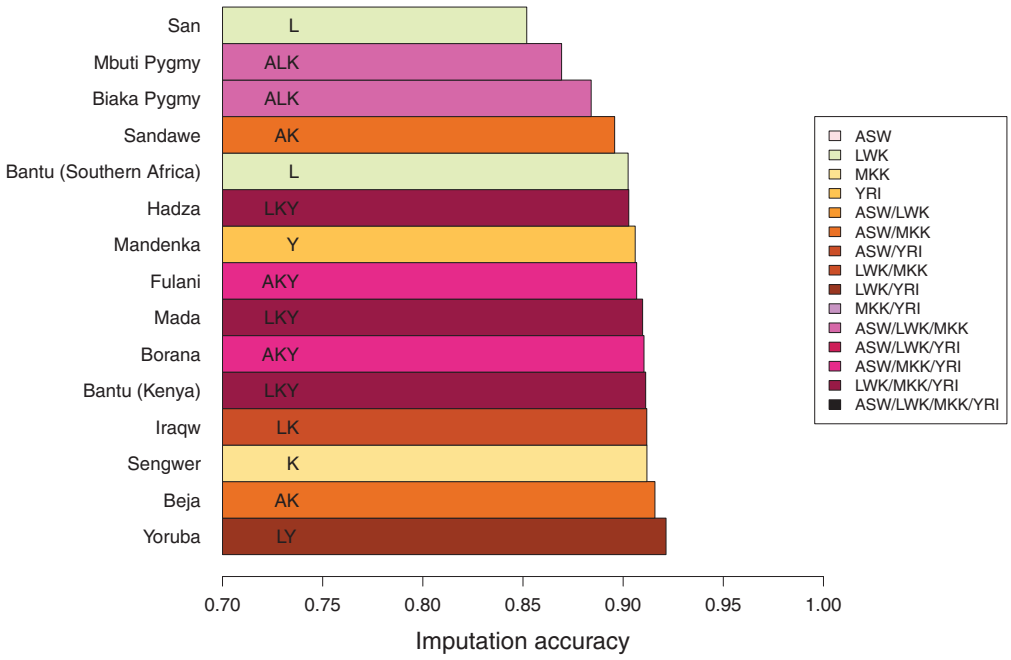

Figure 4.7: Imputation accuracy for inference of genotypes at hidden markers, based on 15 reference panels consisting of combinations among four HapMap Phase 3 panels with recent African ancestry. For each target population, the bar represents the maximal imputation accuracy among the 15 choices, and it is colored according to the choice of optimal reference panel. Each HapMap panel was used with its original size in the combination panels. In each population, we masked the same 77, or $\sim$15%, of 517 markers as in Figure 4.6.

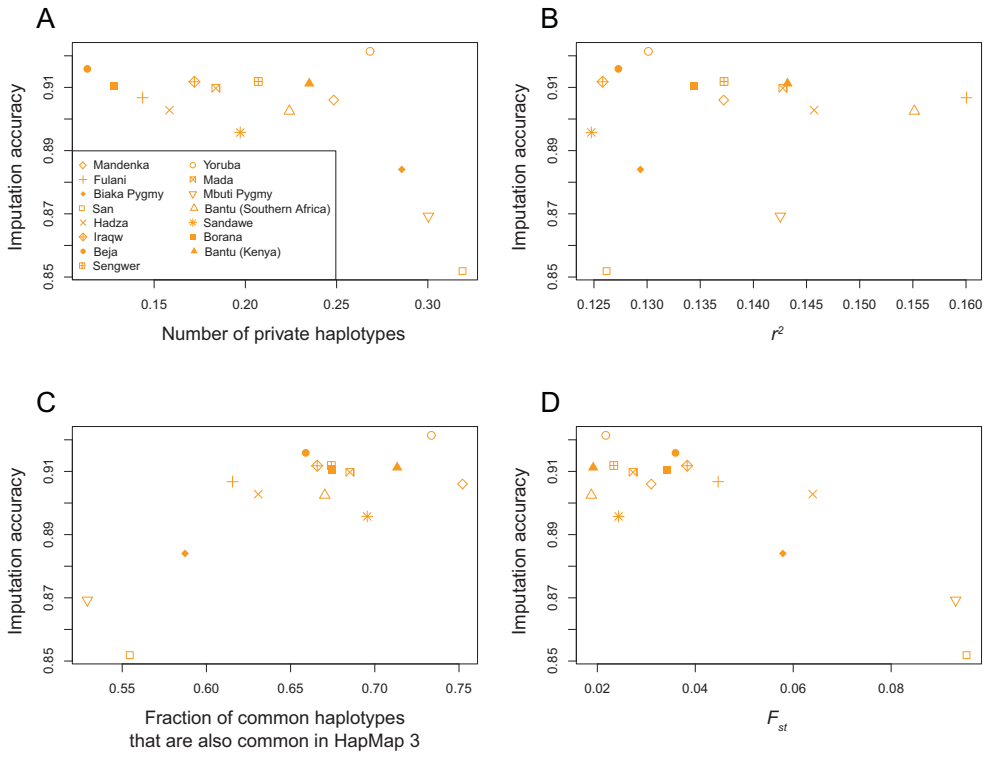

Figure 4.8: Imputation accuracy and statistics of genotypic and haplotypic variation. (A) Number of private haplotypes, (B) linkage disequilibrium as measured by r^2, (C) fraction of common haplotypes also common in the HapMap, and (D) F_{st} between a target population and its optimal HapMap mixture. The imputation accuracy represents the maximal imputation accuracy using the optimal panel among the 15 combinations of the HapMap panels of African descent (identical numerical values as plotted in Figure 4.7). All computations used the set of 517 SNPs that overlapped with HapMap Phase 3. In parts A and C, a window size of 50kb was used; in part B, r^2 was computed using a bin size of 6kb; in part D, F_{st} was first computed for individual SNPs and was then averaged across the 517 SNPs. The fraction of common haplotypes also found in the HapMap and F_{st} were computed for target populations with their respective optimal panels among the 15 choices. The Pearson correlation coefficients are -0.66 ($P = 0.0070$) between imputation accuracy and number of private haplotypes, 0.15 ($P = 0.6044$) between imputation accuracy and r^2, 0.79 ($P = 0.0004$) between imputation accuracy and fraction of common haplotypes in a target population also found in the HapMap, and -0.86 ($P < 0.0001$) between imputation accuracy and F_{st} of a target population with its optimal HapMap mixture.

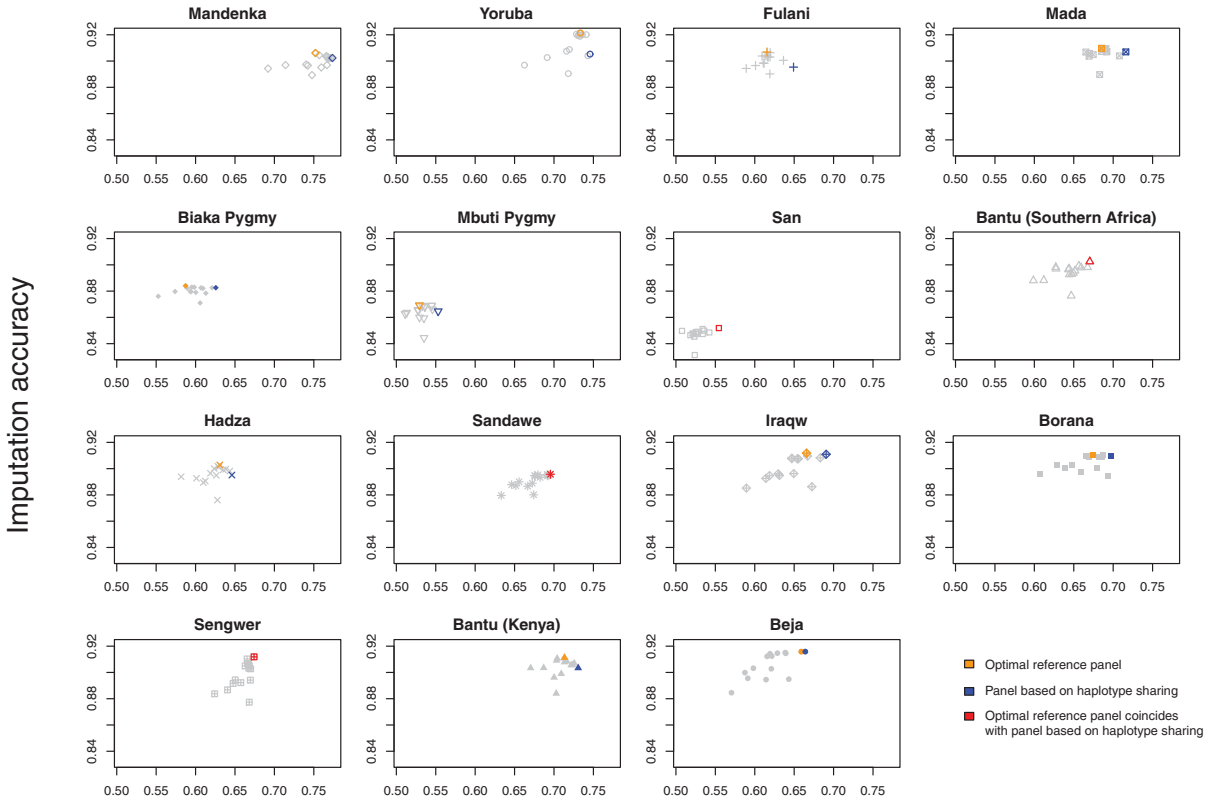

Figure 4.9: Imputation accuracy and the fraction of common haplotypes that are also common in the HapMap. For each target population, imputation accuracy using each of 15 HapMap mixture reference panels is plotted as a function of haplotype sharing with the reference panel. The imputation accuracy for the optimal reference panel corresponds to the maximal imputation accuracy plotted in Figure 4.7.

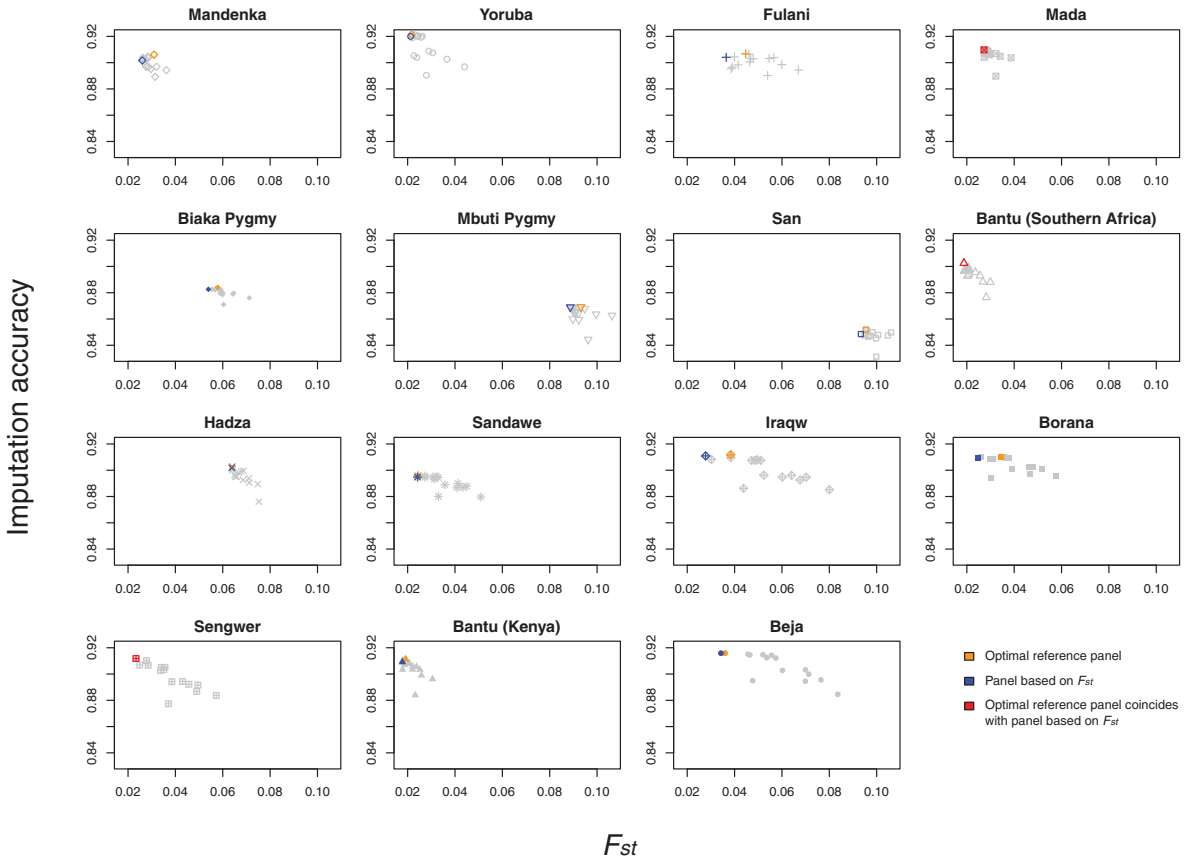

Figure 4.10: Imputation accuracy and F_{st} with HapMap mixtures. For each target population, imputation accuracy using each of 15 HapMap mixture reference panels is plotted as a function of F_{st} with the reference panel. The imputation accuracy for the optimal reference panel corresponds to the maximal imputation accuracy plotted in Figure 4.7.

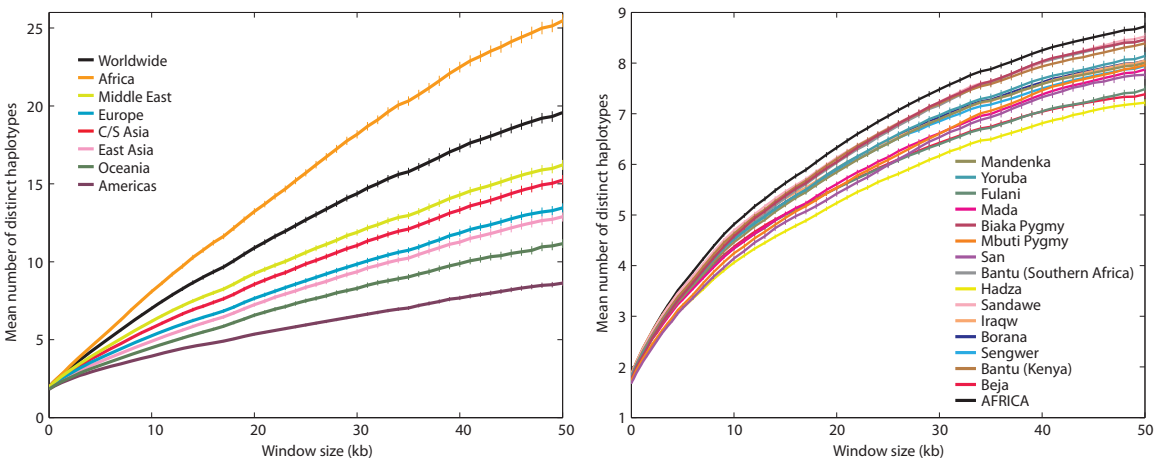

Figure S4.1: Numbers of distinct haplotypes. (A) The number of distinct haplotypes in each geographic region, and in the pooled worldwide collection, as a function of haplotype length. Sample sizes were adjusted to represent 54 chromosomes from each geographic region. (B) The number of distinct haplotypes in each African population, and in the pooled African population, as a function of haplotype length. Sample sizes were adjusted to represent 12 chromosomes from each population. Error bars represent the standard error of the mean across haplotype-loci.

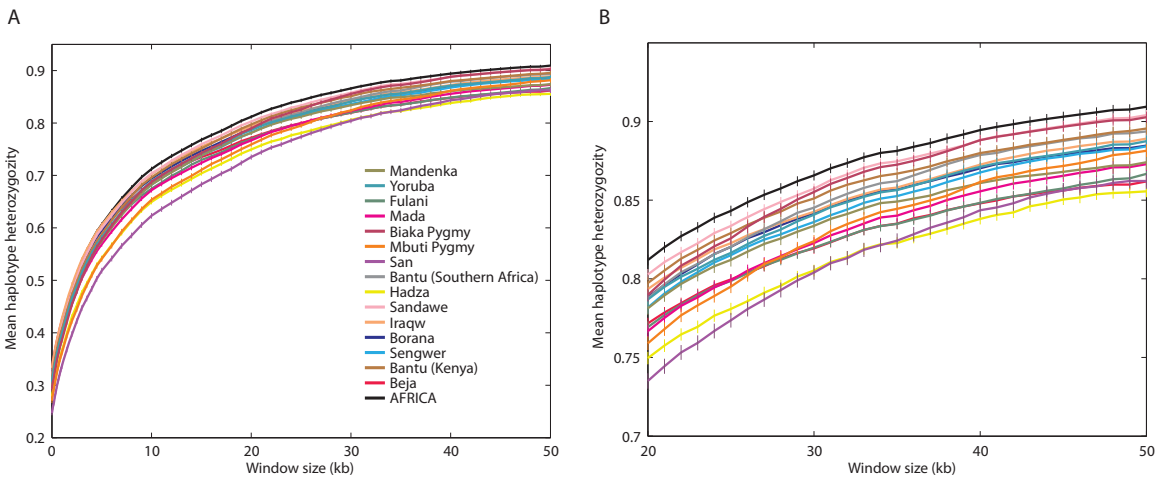

Figure S4.2: Haplotype heterozygosity in African populations. (A) Haplotype heterozygosity in each African population, and in the pooled African population, as a function of haplotype length. Sample sizes were adjusted to represent 12 chromosomes from each population. (B) An amplification of the upper right corner of part A. Error bars represent the standard error of the mean across haplotype-loci.

CHAPTER V

A Coalescent Model for Genotype Imputation

5.1 Introduction

Over the past few years, the field of human genetics has witnessed an explosion in the number of published genome-wide association (GWA) studies, revealing hundreds of novel disease-associated genes (Donnelly, 2008; Manolio *et al.*, 2008; Hindorff *et al.*, 2009, 2011). The considerable potential of GWA studies—which examine thousands to millions of genetic markers in samples of unrelated individuals with the goal of uncovering genotype-phenotype correlations—to ultimately improve human health has been widely recognized (e.g., Pennisi, 2007; Hardy & Singleton, 2009; Manolio, 2010; Stranger *et al.*, 2011).

Among factors contributing to the success of GWA studies is the recent development of genotype-imputation methods (Nicolae, 2006; Li *et al.*, 2006; Browning & Browning, 2007; Marchini *et al.*, 2007; Servin & Stephens, 2007). Genotype imputation is a statistical procedure that uses chromosomal stretches shared among individuals to predict, or impute, genotypes at marker positions not directly measured in individual GWA studies. The haplotypes of "reference" individuals that have been genotyped at a higher density than GWA individuals targeted for imputation often serve as template sequences on the basis of which unknown genotypes in the targets are inferred. Largely because imputation dramatically increases the num-

ber of markers that can be interrogated for disease associations and permits larger sample sizes by enabling data sets typed on different platforms to be merged, it has the potential to increase the statistical power of typical GWA studies (e.g., Li *et al.*, 2009; Marchini & Howie, 2010). This important role for imputation is likely to persist as technology advances. It has been suggested that when whole-genome sequencing of at least a portion of GWA samples becomes feasible for most investigators, imputation will continue to provide a means of improving the power of sequencing-based GWA studies by imputing in genotyped individuals using sequenced individuals as templates (Li *et al.*, 2011).

The determinants of genotype-imputation accuracy have recently been examined empirically in globally distributed human populations (Guan & Stephens, 2008; Pei *et al.*, 2008; Huang *et al.*, 2009a, 2012; Li *et al.*, 2009; Fridley *et al.*, 2010; Surakka *et al.*, 2010). These investigations have collectively shown that in imputation-based GWA studies, population-genetic factors play an important role in determining levels of imputation accuracy attainable in a study population. Factors such as the overall level of linkage disequilibrium (LD) in a study population and the degree of genetic similarity between a study population and a reference population whose members serve as templates have been found in imputation experiments to be prominent drivers of imputation accuracy (Egyud *et al.*, 2009; Huang *et al.*, 2009a, 2012; Paşaniuc *et al.*, 2010; Shriner *et al.*, 2010). However, although this body of empirical work on genotype imputation has provided some understanding of the various population-genetic factors that affect imputation accuracy, analytical work exploring the theoretical basis for the ways in which these factors influence imputation accuracy has been limited.

An analytical approach to studying genotype imputation under a population-genetic model offers the potential for producing a variety of insights. First, by obtaining approximate analytical expressions for the mean and variance of imputation accuracy as a function of population-genetic parameters, we can potentially

explain patterns of imputation accuracy observed in empirical studies in terms of the population-genetic factors that affect the underlying genealogical relationship between study and reference individuals. Second, using simple expressions, imputation accuracy can be evaluated with less computation than in simulation-based approaches, enabling investigators to predict imputation accuracy under a model rather than implement computationally intensive simulations. Third, unlike targeted simulations specific to particular populations of interest, a general population-genetic modeling framework can be adapted for organisms beyond humans in which imputation-based association studies and large-scale genomic resources have begun to emerge (e.g., Atwell *et al.*, 2010; Druet *et al.*, 2010; Kirby *et al.*, 2010).

Jewett *et al.* (2012) recently introduced a theoretical model for evaluating imputation accuracy as a function of population-genetic parameters. Using a coalescent framework, they analytically studied the effect of reference-panel size on imputation accuracy, as well as the degree to which use of reference haplotypes from the same population as a target sequence (an "internal" reference panel) improves the accuracy of imputation compared to use of reference haplotypes from a separate population (an "external" reference panel). In order to incorporate a large sample size in obtaining their analytical results, however, Jewett *et al.* (2012) did not account for randomness in the mutation process. Instead, their treatment of mutation amounted to an assumption that mutation is a deterministic process, in which mutations accumulate along a genealogical branch in direct proportion to the length of the branch. Consequently, under this assumption, the best template for imputation is always a haplotype whose coalescence time with the target sequence on which genotypes are to be imputed is smallest.

Here, we develop a coalescent model with stochastic mutation, and we use this model to explore properties of genotype imputation. Assuming the infinite-sites mutation model, we derive the approximate expectation and variance of imputation

accuracy under a straightforward imputation scheme, conditioning on a mutation parameter (θ), a proportion of markers genotyped in a given length of a chromosome (p), and a time to divergence between the target population and an external reference population (t_d). A distinguishing feature of our approach from that of Jewett *et al.* (2012) is that we explicitly consider a mutation model, thereby allowing for randomness in the imputation process that results from the stochasticity of mutation. As in Jewett *et al.* (2012), in our derivations, we account for randomness in the genealogy by considering the distribution of genealogies under a model in which study and reference individuals are sampled from two populations that diverged at time t_d in the past. Under our model, we pose the following questions: (1) What are the influences of θ, p, and t_d on the expectation and variance of imputation accuracy? (2) What is the expected gain in imputation accuracy in a study sequence targeted for imputation by using a reference sequence from the same population as the target rather than from a different population? Answers to these questions provide information on the factors that affect genotype-imputation accuracy, with implications for the design of imputation-based association studies and the expansion of public genomic databases.

5.2 Theory

In this section, we introduce a theoretical framework that permits the computation of the approximate expectation and variance of imputation accuracy in a target sequence on the basis of a reference sequence. The framework has three parts: a coalescent model that incorporates a mutation process, a decision rule that guides the selection of a reference sequence for the imputation, and an imputation scheme that specifies how the imputation is performed. We compute the expectation and variance of imputation accuracy, conditional on a mutation parameter θ, a proportion p that describes the fraction of sites genotyped in the target sequence, and a population-divergence time t_d.

5.2.1 A Coalescent Model

Consider two populations P_1 and P_2 that diverged from an ancestral population P_A at time t_d in the past. Further, consider three haploid individuals—a study individual targeted for imputation (henceforth simply referred to as a *target* and denoted by I) and two reference individuals (denoted by R_1 and R_2). Reference individual R_1 and target individual I are from population P_1, and reference individual R_2 is from population P_2. In a diploid organism, the haploid individuals can be viewed as single haplotypes.

For the set of three individuals, let $\mathcal{G}$ denote the unobserved random gene tree labeled topology together with a vector $\mathcal{T} = (T_3, T_2, t_d)$ of nonnegative elements that include the unobserved random coalescence times T_k for $k = 3, 2$ and the fixed parametric population-divergence time t_d, where T_k denotes the length of time during which k distinct lineages exist in the genealogy. We assume that the diploid effective population size, denoted by N_e, is the same for the two populations P_1 and P_2 and their ancestral population P_A. We work with coalescent time units so that times are measured in units of $2N_e$ generations. For convenience, we hereafter refer to $\mathcal{G}$ as the genealogy. The genealogy $\mathcal{G}$ can have one of four possible genealogical types G (Figure 5.1): three cases in which the first coalescent event occurs more anciently than the population-divergence time t_d ($g = A, B, C$), and one in which the first coalescent event occurs more recently than time t_d ($g = D$). For each genealogical type, we label the external branches for the lineages of reference individuals R_1 and R_2 and target individual I by 1, 2, and 3, respectively (Figure 5.1).

Throughout this study, we examine the genealogy g backward in time, combining the external branch immediately descended from the root with the internal branch immediately descended from the foot into one branch that takes on the label for the external branch. For instance, branch 2 in genealogy A of Figure 5.1 has length $t_d + t_3 + 2t_2$. Also note that as shown in Figure 5.1, in genealogies A, B and C, T_3

measures from t_d back in time, whereas in genealogy D, T_3 measures from the present and thus includes the length t_d.

Under standard coalescent theory, the time (in units of $2N_e$ generations) for k lineages in the same population to coalesce to $k-1$ lineages follows an exponential distribution with parameter $\binom{k}{2}$ (Wakeley, 2008). Thus, in our model, for $g = A, B, C$ and $k = 3, 2$, and for $g = D$ and $k = 2$,

$$f_{T_k}(t_k; G = g, t_d) = \binom{k}{2} e^{-\binom{k}{2} t_k}, \tag{5.1}$$

where T_k is measured in units of $2N_e$ generations (Figure 5.1). For $g = D$ and $k = 3$, the two lineages in population P_1—reference individual R_1 and target individual I—must coalesce before, or no more anciently than, the divergence time t_d (their ancestral lineage then coalesces with reference individual R_2 in the ancestral population P_A). Hence, given genealogy D and time t_d, the probability density function for the time T_3 to coalescence from three to two lineages is

$$f_{T_3}(t_3; G = D, t_d) = \frac{e^{-t_3}}{1 - e^{-t_d}} \mathbf{1}_{\{t_3 < t_d\}}, \tag{5.2}$$

where T_3 is measured in units of $2N_e$ generations.

The genealogical type G is an unobserved random variable, and we next obtain its probability mass function for use in later computations. More specifically, we compute the probability $\mathbb{P}(G = g | t_d)$ for $g = A, B, C, D$ by conditioning on the location of the coalescence of reference individual R_1 and target individual I. Considering the lineages backward in time, we define $\mathcal{E}$ to be the event that R_1 and I do not coalesce by t_d and $\bar{\mathcal{E}}$ to be the event that R_1 and I do coalesce before t_d. We assume that each pair of lineages in the same population has the same probability of being the first to coalesce. Thus, conditioning on event $\mathcal{E}$, genealogies A, B, and C occur with equal

probabilities. Therefore,

$$\begin{aligned}\mathbb{P}(G=g|t_d) =& \mathbb{P}(G=g|\mathcal{E},t_d)\cdot\mathbb{P}(\mathcal{E}|t_d)+\mathbb{P}(G=g|\bar{\mathcal{E}},t_d)\cdot\mathbb{P}(\bar{\mathcal{E}}|t_d)\\ =& \begin{cases} \frac{1}{3}\cdot e^{-t_d}+0\cdot(1-e^{-t_d})=\frac{1}{3}e^{-t_d} & \text{if } g=A,B,C\\ 0\cdot e^{-t_d}+1\cdot(1-e^{-t_d})=1-e^{-t_d} & \text{if } g=D. \end{cases}\end{aligned} \tag{5.3}$$

Stochastic mutation. In this study, we consider only the polymorphic sites in a sample of three sequences, ignoring all non-polymorphic sites in the sample. We assume the infinite-sites mutation model with no recombination. Under the infinite-sites mutation model, the number of polymorphic sites in a sample is the same as the number of mutations in its gene genealogy, so we use the terms "polymorphic sites" and "mutations" interchangeably. We denote the population-scaled mutation parameter by $\theta = 4N_e\mu L$, where μ is the mutation rate per base pair per generation, and L is the length (in base pairs) of the sequence under consideration. In the remainder of this section, for any genealogical type g, we specify the distributions assumed for random variables that we need later for computing the mean and variance of imputation accuracy.

Let X_i be the unobserved random total number of mutations on branch i under the neutral coalescent model ($i = 1, 2, 3$). We assume that with probability p, a given site is genotyped in the target, and that sites are chosen independently for genotyping. Reference individuals R_1 and R_2 are assumed to be genotyped at all sites at which the set of three lineages is polymorphic. Let Y_i be the random number of mutations on branch i that are genotyped in the target, chosen among all X_i mutations on the branch. Let $h_i(\mathcal{T}; g)$ denote the length of branch i of a given genealogy assumed to have time $\mathcal{T}$ and type g.

The total number of mutations X_i on a branch, conditional on its branch length

$h_i(\mathcal{T}; g)$ follows a Poisson distribution with parameter $h_i(\mathcal{T}; g)\theta/2$. That is,

$$X_i|\mathcal{T}, g \sim \text{Poisson}\left(h_i(\mathcal{T}; g)\theta/2\right), \tag{5.4}$$

where $h_i(\mathcal{T}; g)$ is specified in Table 5.1 for $g = A, B, C, D$ and $i = 1, 2, 3$. Because individual sites are genotyped in the target independently of each other with probability p, conditional on the total number of mutations X_i on branch i, the random number of mutations Y_i on branch i that are genotyped in the target follows a binomial distribution with parameters X_i and p,

$$Y_i|X_i \sim \text{Bin}(X_i, p). \tag{5.5}$$

The unobserved random number of mutations $(X_i - Y_i)$ on branch i that are untyped in the target follows a binomial distribution with parameters X_i and $1 - p$,

$$(X_i - Y_i)|X_i \sim \text{Bin}(X_i, 1 - p). \tag{5.6}$$

Equations 5.4 and 5.5 imply that

$$Y_i|\mathcal{T}, g \sim \text{Poisson}(h_i(\mathcal{T}; g)\theta p/2) \tag{5.7}$$

(Casella & Berger, 2001, pg. 163). Similarly, equations 5.4 and 5.6 imply that conditional on the coalescence and population-divergence times $\mathcal{T}$, the number of mutations $(X_i - Y_i)$ on branch i that are untyped in the target follows a Poisson distribution with parameter $h_i(\mathcal{T}; g)\theta(1 - p)/2$,

$$(X_i - Y_i)|\mathcal{T}, g \sim \text{Poisson}(h_i(\mathcal{T}; g)\theta(1 - p)/2). \tag{5.8}$$

Furthermore, conditional on $\mathcal{T}$ the numbers of mutations on any two branches Y_i and

Y_j that are genotyped in the target are independent, and the difference in the two independent Poisson-distributed variables follows a Skellam distribution (Johnson & Kotz, 1969). Thus, for $i, j = 1, 2, 3$ and $i \neq j$,

$$(Y_i - Y_j)|\mathcal{T}, g \sim \text{Skellam}(h_i(\mathcal{T}; g)\theta p/2, h_j(\mathcal{T}; g)\theta p/2), \tag{5.9}$$

with mean $(h_i(\mathcal{T}; g) - h_j(\mathcal{T}; g))\theta p/2$ and variance $(h_i(\mathcal{T}; g) + h_j(\mathcal{T}; g))\theta p/2$.

5.2.2 A Decision Rule

Recall that in our sample of three haploid individuals—a target I and two references R_1 and R_2, we consider only polymorphic sites. The target individual is assumed to be genotyped at only a subset of the sites that are polymorphic in the three individuals. We now further assume that missing genotypes at untyped markers in the target are substituted and thus imputed with corresponding genotypes in a chosen reference individual who has been genotyped at all of the sites. The choice of a reference individual for the imputation is specified by a decision rule δ_s that we introduce below, and the metric that we use for evaluating imputation accuracy is specified in Section 5.2.3.

Generally, because imputation relies on the occurrence of chromosomal stretches that are shared identically by descent between target and reference individuals, we expect imputation accuracy in a target individual to improve with increased genetic similarity between the target and reference individuals. We therefore define a distance statistic d_i between reference individual R_i (branch i; $i = 1, 2$) and target individual I (branch 3) to be the number of pairwise sequence differences between the two individuals at positions genotyped in the target, namely

$$d_i := Y_i + Y_3. \tag{5.10}$$

Smaller values of d_i indicate a greater degree of observed genetic similarity—measured at the genotyped positions in the target—between reference individual R_i and the target.

We now present a decision rule (δ_s) based on the distance statistic d_i (eq. 5.10) that we use to select which of the two reference individuals, R_1 or R_2, is used for imputation in the target individual. In short, choosing between the two reference individuals, rule δ_s selects the genetically more similar reference individual to the target as measured by the distance statistic (i.e., by the observed number of pairwise sequence differences between the reference and the target).

Rule δ_s

- If $d_1 < d_2$, use reference R_1.
- If $d_2 < d_1$, use reference R_2.
- If $d_1 = d_2$, with probability 1/2, use reference R_1, and with probability 1/2, use reference R_2.

5.2.3 An Imputation Scheme

Once a reference individual is chosen for imputation in the target, we substitute missing genotypes at untyped markers in the target by those at corresponding positions in the reference. We illustrate the reference selection and the imputation procedure in Figure 5.2.

We assess imputation accuracy by the proportion of polymorphic sites untyped in the target that are subsequently imputed correctly on the basis of a chosen reference individual. Let R_i, $i = 1, 2$, denote the chosen reference individual, and let R_j, $j \neq i$ and $j = 2, 1$, denote the reference individual that is not chosen. Then, imputation accuracy obtained on the basis of reference individual R_i is defined as

$$Z := \frac{X_j - Y_j}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}. \tag{5.11}$$

We note that the denominator $\sum_{\ell=1}^{3}(X_\ell - Y_\ell)$ corresponds to the total number of untyped polymorphic sites in the target that are subsequently imputed, and that when $\sum_{\ell=1}^{3}(X_\ell - Y_\ell) = 0$, there are no genotypes to impute, rendering Z undefined. The numerator is $X_j - Y_j$ because under the infinite-sites mutation model, among polymorphic sites, those produced by mutations on the branch corresponding to reference individual R_j are exactly where reference individual R_i and target individual I have identical genotypes. Thus, to count the number of polymorphic sites imputed correctly in the target on the basis of reference individual R_i, one simply counts the number of mutations on the branch corresponding to reference individual R_j that are not genotyped in the target but that are imputed.

5.2.4 Approximate Expressions for the Expectation and Variance of Imputation Accuracy

At sites genotyped in both reference and target individuals, the number of pairwise sequence differences d_i between reference individual R_i $(i = 1, 2)$ and target individual I is known. Given d_i for $i = 1, 2$, we apply the rule δ_s in Section 5.2.2 to select a reference individual for imputing missing genotypes at untyped markers in the target. Conditioning on the model parameters—the mutation parameter θ, the proportion p of polymorphic markers genotyped in the target, and the population-divergence time t_d—in this section, we derive the approximate expectation and variance of imputation accuracy Z defined in eq. 5.11 by averaging over all possible genealogical types G and coalescence times T_3 and T_2.

To compute the expectation $\mathbb{E}[Z|\theta, p, t_d]$, we consider three possible scenarios that can occur when we apply rule δ_s to a genealogy: reference individual R_1 is selected as the template sequence for imputation in target individual I because $d_1 < d_2$, reference individual R_2 is selected because $d_1 > d_2$, and a choice is made probabilistically between references R_1 and R_2 because $d_1 = d_2$. Let $\mathcal{S}_1$ be the scenario in which $d_1 < d_2$

(i.e., $Y_1 - Y_2 < 0$), let $\mathcal{S}_2$ be the scenario in which $d_1 > d_2$ (i.e., $Y_1 - Y_2 > 0$), and let $\mathcal{S}_3$ be the scenario in which $d_1 = d_2$ (i.e., $Y_1 - Y_2 = 0$). We can obtain $\mathbb{E}[Z|\theta, p, t_d]$ by taking a weighted average of its expectation conditional on the genealogical type g and the scenario $\mathcal{S}_x$, where $g = A, B, C, D$ and $x = 1, 2, 3$, and where the weight is the joint probability of the genealogical type $G = g$ and the scenario $\mathcal{S}_x$:

$$\mathbb{E}[Z|\theta, p, t_d] \;=\; \sum_{g=A,B,C,D} \sum_{x=1}^{3} \mathbb{E}[Z|g, \mathcal{S}_x, \theta, p, t_d]\mathbb{P}(g, \mathcal{S}_x|\theta, p, t_d). \tag{5.12}$$

We first derive the conditional expectations $\mathbb{E}[Z|g, \mathcal{S}_x, \theta, p, t_d]$ and the probabilities $\mathbb{P}(g, \mathcal{S}_x|\theta, p, t_d)$ for $g = A, B, C, D$ and $x = 1, 2, 3$, and we then obtain the expectation $\mathbb{E}[Z|\theta, p, t_d]$ using eq. 5.12.

All quantities in Sections 5.2.4.1 and 5.2.4.2 are conditional on θ, p and t_d, but for notational convenience, these parameters are suppressed.

5.2.4.1 Derivation of $\mathbb{E}[Z|g, \mathcal{S}_x]$ in eq. 5.12

Let B be a Bernoulli random variable with parameter 1/2. For any genealogical type g, we can explicitly express the expectation $\mathbb{E}[Z|g, \mathcal{S}_x]$ under a specific scenario $\mathcal{S}_x$ for $x = 1, 2, 3$ as follows:

$$\mathbb{E}[Z|g, \mathcal{S}_x] = \begin{cases} \mathbb{E}\left[\frac{X_j - Y_j}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\middle| g, \mathcal{S}_x\right] \text{ for } x, j = 1, 2 \text{ and } j \neq x \\ \mathbb{E}\left[b\frac{X_2 - Y_2}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)} + (1-b)\frac{X_1 - Y_1}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\middle| g, \mathcal{S}_x\right] \text{ for } x = 3. \end{cases} \tag{5.13}$$

For computational convenience, to obtain $\mathbb{E}[Z|g, \mathcal{S}_x]$, we use the first-order Taylor-series approximation and an additional approximation of $\mathbb{E}[X_j - Y_j|g, \mathcal{S}_x]$ by $\mathbb{E}[X_j - Y_j|g]$. That is, for $g = A, B, C, D$, $x, j = 1, 2$, and $j \neq x$,

$$\mathbb{E}[Z|g, \mathcal{S}_x] = \mathbb{E}\left[\frac{X_j - Y_j}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\middle| g, \mathcal{S}_x\right]$$

$$\approx \frac{\mathbb{E}[X_j - Y_j | g, \mathcal{S}_x]}{\mathbb{E}[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g, \mathcal{S}_x]} \tag{5.14}$$

$$\approx \frac{\mathbb{E}[X_j - Y_j | g]}{\mathbb{E}[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g]}, \tag{5.15}$$

and for $g = A, B, C, D$ and $x = 3$,

$$\mathbb{E}[Z|g, \mathcal{S}_x] = \frac{1}{2}\left(\mathbb{E}\left[\frac{X_2 - Y_2}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\middle| g, \mathcal{S}_x\right] + \mathbb{E}\left[\frac{X_1 - Y_1}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\middle| g, \mathcal{S}_x\right]\right) \tag{5.16}$$

$$\approx \frac{1}{2}\left(\frac{\mathbb{E}[X_2 - Y_2|g, \mathcal{S}_x]}{\mathbb{E}[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g, \mathcal{S}_x]} + \frac{\mathbb{E}[X_1 - Y_1|g, \mathcal{S}_x]}{\mathbb{E}[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g, \mathcal{S}_x]}\right) \tag{5.17}$$

$$\approx \frac{1}{2}\left(\frac{\mathbb{E}[X_2 - Y_2|g]}{\mathbb{E}[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g]} + \frac{\mathbb{E}[X_1 - Y_1|g]}{\mathbb{E}[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g]}\right). \tag{5.18}$$

In eqs. 5.15 and 5.18, for $g = A, B, C, D$ and $i = 1, 2, 3$, the expectation $\mathbb{E}[X_i - Y_i|g]$ can be found by conditioning on the coalescence times T_3 and T_2 and then integrating over their distributions:

$$\mathbb{E}[X_i - Y_i|g] = \int_{t_2=0}^{\infty}\int_{t_3=0}^{a} \mathbb{E}[X_i - Y_i|t_3, t_2, g] \cdot f_{T_3,T_2}(t_3, t_2|g)\, \mathrm{d}t_3\, \mathrm{d}t_2. \tag{5.19}$$

As the expectation of a Poisson random variable,

$$\mathbb{E}[X_i - Y_i|t_3, t_2, g] = h_i(\mathcal{T};g)\theta(1-p)/2. \tag{5.20}$$

In any genealogy, by the independence of coalescence times under the coalescent model,

$$f_{T_3,T_2}(t_3, t_2|g) = f_{T_3}(t_3|g) \cdot f_{T_2}(t_2|g), \tag{5.21}$$

where $f_{T_3}(t_3|g)$ and $f_{T_2}(t_2|g)$ are evaluated using eqs. 5.1 and 5.2. The upper limit of

the inner integral in eq. 5.19 depends on the genealogy under consideration; that is,

$$a = \begin{cases} \infty, & \text{if } g = A, B, C \\ t_d, & \text{if } g = D. \end{cases} \tag{5.22}$$

This completes the derivation of $\mathbb{E}[Z|g, \mathcal{S}_x]$ in eq. 5.12, and we now derive $\mathbb{P}(g, \mathcal{S}_x)$.

5.2.4.2 Derivation of $\mathbb{P}(g, \mathcal{S}_x)$ in eq. 5.12

We compute the probability $\mathbb{P}(g, \mathcal{S}_x)$ by jointly considering the marginal distribution of g and the conditional distribution of $\mathcal{S}_x$ given a genealogical type $G = g$:

$$\mathbb{P}(g, \mathcal{S}_x) = \mathbb{P}(g) \cdot \mathbb{P}(\mathcal{S}_x|g). \tag{5.23}$$

The probability $\mathbb{P}(g)$ is given in eq. 5.3. As in the derivation of the expectation $\mathbb{E}[X_i - Y_i|g]$ in eq. 5.19, to compute $\mathbb{P}(\mathcal{S}_x|g)$, we first condition on the coalescence times T_3 and T_2 and then integrate over their distributions:

$$\mathbb{P}(\mathcal{S}_x|g) = \int_{t_2=0}^{\infty} \int_{t_3=0}^{a} \mathbb{P}(\mathcal{S}_x|t_3, t_2, g) \cdot f_{T_3,T_2}(t_3, t_2|g) \, \mathrm{d}t_3 \, \mathrm{d}t_2, \tag{5.24}$$

where a is given in eq. 5.22, $\mathbb{P}(\mathcal{S}_x|t_3, t_2, g)$ can be obtained by considering the difference $Y_1 - Y_2$ and using eq. 5.9, and $f_{T_3,T_2}(t_3, t_2|g)$ can be computed using eqs. 5.1, 5.2, and 5.21.

This completes the derivation of the expectation $\mathbb{E}[Z]$ in eq. 5.12.

5.2.4.3 Derivation of $\mathrm{Var}[Z|\theta, p, t_d]$

We first note that we can obtain $\mathrm{Var}[Z|\theta, p, t_d]$ using the following equation:

$$\mathrm{Var}[Z|\theta, p, t_d] = \mathbb{E}[Z^2|\theta, p, t_d] - \mathbb{E}[Z|\theta, p, t_d]^2, \tag{5.25}$$

where $\mathbb{E}[Z|\theta, p, t_d]$ is as derived above (eq. 5.12). It remains to derive $\mathbb{E}[Z^2|\theta, p, t_d]$.

As in the derivation of the expectation $\mathbb{E}[Z|\theta, p, t_d]$ (eq. 5.12), we obtain the expectation $\mathbb{E}[Z^2|\theta, p, t_d]$ by conditioning on the genealogical type g and the scenario $\mathcal{S}_x$, where $g = A, B, C, D$ and $x = 1, 2, 3$:

$$\mathbb{E}[Z^2|\theta, p, t_d] \;=\; \sum_{g=A,B,C,D} \sum_{x=1}^{3} \mathbb{E}[Z^2|g, \mathcal{S}_x, \theta, p, t_d]\mathbb{P}(g, \mathcal{S}_x|\theta, p, t_d). \qquad (5.26)$$

Because the derivation of $\mathbb{E}[Z^2|\theta, p, t_d]$ (eq. 5.26) is similar to that of $\mathbb{E}[Z|\theta, p, t_d]$ (eq. 5.12), we omit it to avoid redundancy. The main difference in the derivation here is the first-order Taylor series approximation for the variance $\mathrm{Var}[Z|g, \mathcal{S}_x, \theta, p, t_d]$ needed in estimating the expectation $\mathbb{E}[Z^2|g, \mathcal{S}_x, \theta, p, t_d]$ in eq. 5.26:

$$\begin{aligned}
\mathrm{Var}[Z|g, \mathcal{S}_x, \theta, p, t_d] =& \mathrm{Var}\left[\frac{X_j - Y_j}{\sum_{\ell=1}^{3}(X_\ell - Y_\ell)}\middle| g, \mathcal{S}_x, \theta, p, t_d\right] \\
\approx& \left(\frac{\mathbb{E}[X_j - Y_j|g, \mathcal{S}_x, \theta, p, t_d]}{\mathbb{E}[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g, \mathcal{S}_x, \theta, p, t_d]}\right)^2 \left(\frac{\mathrm{Var}[X_j - Y_j|g, \mathcal{S}_x, \theta, p, t_d]}{\mathbb{E}[X_j - Y_j|g, \mathcal{S}_x, \theta, p, t_d]^2}\right. \\
&+ \frac{\mathrm{Var}[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g, \mathcal{S}_x, \theta, p, t_d]}{\mathbb{E}[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g, \mathcal{S}_x, \theta, p, t_d]^2} \\
&\left. - \frac{2\mathrm{Cov}(X_j - Y_j, \sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g, \mathcal{S}_x, \theta, p, t_d)}{\mathbb{E}[X_j - Y_j|g, \mathcal{S}_x, \theta, p, t_d]\mathbb{E}[\sum_{\ell=1}^{3}(X_\ell - Y_\ell)|g, \mathcal{S}_x, \theta, p, t_d]}\right)
\end{aligned}$$

(Casella & Berger, 2001, pg. 245). The quantity $\mathbb{P}(g, \mathcal{S}_x|\theta, p, t_d)$ in eq. 5.26 is obtained as in Section 5.2.4.2.

5.3 Methods of Computation and Simulation

To calculate the expectation $\mathbb{E}[Z|\theta, p, t_d]$ (eq. 5.12) in practice, we obtain approximations for $\mathbb{E}[Z|g, \mathcal{S}_x, \theta, p, t_d]$ using eqs. 5.15 and 5.18 and the Monte Carlo estimates of $\mathbb{P}(\mathcal{S}_x|g, \theta, p, t_d)$ involved in the expression of $\mathbb{P}(g, \mathcal{S}_x|\theta, p, t_d)$. To compute the variance $\mathrm{Var}[Z|\theta, p, t_d]$ (eq. 5.25), we further approximate $\mathbb{E}[Z^2|g, \mathcal{S}_x, \theta, p, t_d]$ in eq. 5.26. Finally, to verify the approximate analytical expressions of $\mathbb{E}[Z|\theta, p, t_d]$ and

$\text{Var}[Z|\theta, p, t_d]$ in eqs. 5.12 and 5.25, using the algorithm below, we performed simulations to obtain simulated means and variances of imputation accuracy that we then compared to our estimates of $\mathbb{E}[Z|\theta, p, t_d]$ and $\text{Var}[Z|\theta, p, t_d]$.

Algorithm for estimating $\mathbb{E}[Z|\theta, p, t_d]$ ***and*** $\text{Var}[Z|\theta, p, t_d]$ ***through simulation***

1. Set parameter values for θ, p, and t_d.
2. For $m = 1$ to M:
   (a) Generate a genealogical type G using a uniformly distributed random variable $U \sim \text{Uniform}(0, 1)$. If $u < 1 - e^{-t_d}$, set $g = D$. Otherwise, generate u' from $\text{Uniform}(0, 1)$, independently of u. Set $g = A$ if $u' \in [0, 1/3)$, set $g = B$ if $u' \in [1/3, 2/3)$, and set $g = C$ if $u' \in [2/3, 1)$.
   (b) Generate a coalescence time $T_2 \sim \text{Exp}(1)$.
   (c) If $g = A, B, C$, generate a coalescence time $T_3 \sim \text{Exp}(3)$. Otherwise, generate T_3 from the probability density function in eq. 5.2.
   (d) For $i = 1, 2, 3$, generate a total number of mutations $X_i \sim \text{Poisson}(h_i(\mathcal{T}; g)\theta/2)$ on branch i, where $\mathcal{T} = (t_3, t_2, t_d)$ and $h_i(\mathcal{T}; g)$ is specified in Table 5.1.
   (e) For $i = 1, 2, 3$, given X_i, sample the number of mutations on branch i that are genotyped in the target as $Y_i|X_i = x_i \sim \text{Binomial}(x_i, p)$.
   (f) If $\sum_{i=1}^{3}(x_i - y_i) = 0$, return to (b); otherwise, continue.
   (g) If $y_1 - y_2 < 0$ (i.e., if $d_1 < d_2$), compute $z_{(m)} = \frac{x_2 - y_2}{\sum_{i=1}^{3}(x_i - y_i)}$.
   (h) If $y_1 - y_2 > 0$ (i.e., if $d_2 < d_1$), compute $z_{(m)} = \frac{x_1 - y_1}{\sum_{i=1}^{3}(x_i - y_i)}$.
   (i) If $y_1 - y_2 = 0$ (i.e., if $d_1 = d_2$), generate b from Bernoulli(1/2). Compute $z_{(m)} = \frac{x_2 - y_2}{\sum_{i=1}^{3}(x_i - y_i)}$ if $b = 1$ and $z_{(m)} = \frac{x_1 - y_1}{\sum_{i=1}^{3}(x_i - y_i)}$ otherwise.
3. Compute the sample mean $\bar{z}$ and the sample variance s^2 that respectively represent simulation-based estimates of $\mathbb{E}[Z|\theta, p, t_d]$ and $\text{Var}[Z|\theta, p, t_d]$, where $\bar{z} = \frac{1}{M}\sum_{m=1}^{M} z_{(m)}$ and $s^2 = \frac{1}{M-1}\sum_{m=1}^{M}(z_{(m)} - \bar{z})^2$.

5.4 The Role of the Parameters

Figures 5.3A and 5.3B separately plot the expected imputation accuracy $\mathbb{E}[Z|\theta, p, t_d]$ as a function of the mutation rate θ and the proportion p of polymorphic sites that are genotyped in a target sequence. For the parameter values considered, the theoretical approximations of $\mathbb{E}[Z|\theta, p, t_d]$ obtained using eq. 5.12 closely match the simulated averages of imputation accuracy, although the theoretical approach tends to produce smaller values than the simulation for most parameter values considered in Figure 5.3.

The observed underestimation of the theoretical estimates of $\mathbb{E}[Z|\theta, p, t_d]$ by eq. 5.12 can possibly be attributed to the Taylor series approximation for the expectation $\mathbb{E}[Z|g, \mathcal{S}_x, \theta, p, t_d]$ (eqs. 5.14 and 5.17) and to the additional approximation of $\mathbb{E}[X_i - Y_i|g, \mathcal{S}_x, \theta, p, t_d]$ by $\mathbb{E}[X_i - Y_i|g, \theta, p, t_d]$ in estimating $\mathbb{E}[Z|g, \mathcal{S}_x, \theta, p, t_d]$ (eqs. 5.15 and 5.18). However, at $p = 0$ when only the Taylor series approximation has an effect, there is no evidence of such underestimation (Figure 5.3A), suggesting that the underestimation is primarily due to the additional approximation $\mathbb{E}[X_j - Y_j|g, \mathcal{S}_x, \theta, p, t_d] \approx \mathbb{E}[X_j - Y_j|g, \theta, p, t_d]$. In this additional approximation, for ease of computation, we ignore the condition $\mathcal{S}_x$ in calculating the expected values of the number of sites correctly imputed in the target, $X_j - Y_j$ (i.e., the numerator of the imputation-accuracy statistic Z). However, $\mathcal{S}_x$ provides information about which reference individual has fewer pairwise sequence differences with the target. Therefore, knowing $\mathcal{S}_x$ leads one to expect higher imputation accuracy than not knowing the condition, and disregarding $\mathcal{S}_x$ in our approximation may be partially responsible for the observed underestimation.

The mutation parameter θ plays an important role in determining imputation accuracy in a target sequence, as long as the proportion of polymorphic sites that are genotyped in the target p is non-zero. More specifically, for $p > 0$, expected imputation accuracy increases as θ increases, whereas for $p = 0$, expected imputation

accuracy stays constant as θ increases (Figure 5.3A). When $p > 0$, increasing θ increases the number of pairwise sequence differences between each reference sequence and the target. The larger sequence differences in turn enable a more accurate determination of which reference sequence is genetically more similar to the target.

When p is small, compared to the case in which p is large, the expected imputation accuracy increases more steadily as θ increases from 1 to 10 (Figure 5.3A). Consider the total increase in the expected imputation accuracy as θ increases from 1 to 10. This increase is comparable for $p = 0.1$ and $p = 0.9$ (0.0858 and 0.0877, respectively); while fifty percent of such increase do not occur until $\theta \in (4, 5)$ for $p = 0.1$, they occur by $\theta \in (2, 3)$ for $p = 0.9$. This result suggests that at small values of θ, the rate of increase in the expected imputation accuracy due to an increase in θ is slower for a small, rather than large, value of p. In other words, when the proportion p of polymorphic sites that are genotyped in the target is small, the effect of the mutation parameter θ on imputation accuracy is less pronounced than when p is large.

The proportion p of polymorphic sites that are genotyped in a target sequence also plays an important role in determining imputation accuracy in the target. As p increases, imputation accuracy increases in expectation (Figure 5.3B). We again offer the intuitive explanation that increased p implies increased information for selecting a reference sequence that has an increased probability of being more genuinely similar to the target. For each θ that we considered, the rate of increase in the expected imputation accuracy due to an increase in p is faster when p is small than when p is large. This result is consistent with empirical observations of Fridley *et al.* (2010), who found that in a candidate gene sequencing study, imputation accuracy improved with an increase in the number of markers genotyped in the target and that the improvement was most pronounced when the initial number of genotyped markers was smallest.

Figures 5.4A and 5.4B separately plot the variance in imputation accuracy $\mathrm{Var}[Z|\theta, p, t_d]$

as a function of θ and p. Except when θ is small (e.g., $\theta \leq 4$) or when p is large (e.g., $p > 0.5$), the theoretical estimates of $\text{Var}[Z|\theta, p, t_d]$ obtained using eq. 5.25 closely match the simulated variances. Further, we find that the variance of imputation accuracy increases with decreasing θ and with increasing p. This increase occurs because decreasing θ and increasing p reduce the number of polymorphic sites that need to be imputed in the target, thereby producing a larger variance in the statistic measuring imputation accuracy.

For two values of the divergence time t_d between the target and reference populations ($t_d = 0$ and $t_d = 0.1$), Figures 5.5A and 5.5B separately plot the approximate expected imputation accuracy $\mathbb{E}[Z|\theta, p, t_d]$ as a function of θ and p. For both values of t_d, expected imputation accuracy increases with increasing θ and p. Moreover, for given values of θ and p, expected imputation accuracy also increases as t_d increases. This is because with increasing t_d, there is more time for mutations to accumulate along each branch, on average. For fixed values of θ and p, an increase in the occurrence of mutations leads to an increase in the number of polymorphic sites that are genotyped in the target, again allowing a more accurate determination of the reference sequence that is genetically more similar to the target.

For $t_d = 0$ and $t_d = 0.1$, Figures 5.6A and 5.6B separately plot the approximate variance of imputation accuracy $\text{Var}[Z|\theta, p, t_d]$ as a function of θ and p. For both values of t_d, we observe the same patterns of increasing variance in imputation accuracy with decreasing θ and increasing p.

To assess the expected gain in imputation accuracy in a target sequence by using a reference sequence from the same population as the target (i.e., reference sequence R_1) rather than from a different population (i.e., reference sequence R_2), for $t_d = 0.1$ and $t_d = 0.01$, we plotted the simulated mean imputation accuracies in the target for each scenario separately in Figure 5.7A. The simulation procedure was modified to produce the results in both plots of Figure 5.7, with steps (g)-(i) replaced by a

single step of computing $z_{(m)} = \frac{x_2 - y_2}{\sum_{i=1}^{3}(x_i - y_i)}$ for the case in which R_1 is always used as the template and $z_{(m)} = \frac{x_1 - y_1}{\sum_{i=1}^{3}(x_i - y_i)}$ otherwise. The difference in the mean accuracies between the imputations performed using R_1 and the imputations performed using R_2 remains fairly constant over a wide range of parameter values considered for θ and p (Figure 5.7B). Denote the mean value of this difference by Δ_{t_d}. Then, for $\theta = 1, 2, \cdots, 10$ and $p = 0, 0.1, 0.5, 0.9$, $\Delta_{0.1} = 0.0829$ and $\Delta_{0.01} = 0.0099$ based on simulations. Surprisingly, we can estimate these mean differences quite accurately using a simple formula:

$$\widehat{\Delta}_{t_d} = \frac{2\mathbb{E}[T_2|G = D, t_d] + 2t_d - 2\mathbb{E}[T_3|G = D, t_d]}{2\mathbb{E}[T_2|G = D, t_d] + 2t_d + \mathbb{E}[T_3|G = D, t_d]}\mathbb{P}(G = D|t_d), \tag{5.27}$$

where $\mathbb{E}[T_2|G = D, t_d]$ and $\mathbb{E}[T_3|G = D, t_d]$ are found using eqs. 5.1 and 5.2, respectively, and where $\mathbb{P}(G = D|t_d) = 1 - e^{-t_d}$ is given in eq. 5.3. We obtain that for any $t_d \geq 0$,

$$\mathbb{E}[T_2|G = D, t_d] = 1 \tag{5.28}$$

and

$$\mathbb{E}[T_3|G = D, t_d] = \int_0^{t_d} t_3 \frac{e^{-t_3}}{1 - e^{-t_d}}\, \mathrm{d}t_3 = \frac{1 - (1 + t_d)e^{-t_d}}{1 - e^{-t_d}}. \tag{5.29}$$

Evaluating all the terms involved in $\widehat{\Delta}_{t_d}$ (eq. 5.27) at $t_d = 0.1$ and $t_d = 0.01$, we have $\widehat{\Delta}_{0.1} = 0.0889$ and $\widehat{\Delta}_{0.01} = 0.0099$.

To show that use of reference sequence R_1 always results in higher imputation accuracy in the target, on average, than use of reference sequence R_2, we prove here $\widehat{\Delta}_{t_d} \geq 0$ by noting that the numerator and denominator of $\widehat{\Delta}_{t_d}$ (eq. 5.27) are positive numbers. This is because in genealogy D, times T_2, T_3, and t_d must be non-negative, and T_3 must be smaller than or equal to t_d (see Figure 5.1). Therefore, in eq. 5.27,

the numerator

$$2\{\mathbb{E}[T_2|G = D, t_d] + (t_d - \mathbb{E}[T_3|G = D, t_d])\} > 2\{0 + 0\} > 0,$$

and the denominator

$$2\{\mathbb{E}[T_2|G = D, t_d] + t_d\} + \mathbb{E}[T_3|G = D, t_d] > 2\{0 + 0\} + 0 > 0.$$

5.5 Discussion

We have introduced a theoretical framework for investigating genotype imputation and the various population-genetic factors that affect imputation accuracy. The framework includes a two-population coalescent model for three sequences, and a mutation model to account for stochasticity in the mutation process and thus in the choice of imputation template. Using this framework and a simple imputation scheme, we have derived approximate expressions for the expectation and variance of the accuracy of imputation in the target sequence using a reference sequence chosen on the basis of observed genetic similarity to the target at genotyped positions.

The three parameters of the coalescent model are the mutation parameter θ, the proportion p of polymorphic sites in a chromosomal region that are genotyped in the target, and the divergence time t_d between the two populations. Measuring imputation accuracy by the proportion of polymorphic sites that are untyped but subsequently imputed correctly in the target, we found that imputation accuracy increases in expectation with increasing θ, p, and t_d. We also observed that the variance in imputation accuracy decreases with increasing θ and t_d, and that the variance increases with increasing p. Additionally, we found that under the model, the expected gain in

accuracy when the reference sequence R_1, rather than the reference sequence R_2, is used can be accurately predicted by a simple formula relating the expected difference in imputation accuracy to the expected difference in branch lengths of R_1 and R_2 (eq. 5.27).

Our results on the trends in the expected imputation accuracy can be explained intuitively by considering the amount of information available for determining which of the two reference sequences, R_1 or R_2, is genetically more similar to the target. For instance, increasing the mutation parameter $\theta = 4N_e\mu L$ can be considered equivalent to increasing the length L of the chromosomal region under consideration, when the effective population size N_e and the mutation rate μ per base pair per generation are held constant. For any p and t_d, comparing genotypes at a fixed proportion p of markers in longer target and reference sequences rather than in shorter ones increases the probability that the reference sequence that is truly genetically more similar to the target at typed and untyped markers alike in the region of interest can be identified. An increase in the probability of correctly identifying the genetically more similar reference sequence, and therefore using it for imputation, leads to an increase in expected imputation accuracy.

We conclude with a discussion of model limitations. Because of the complexity in computing the expectation and variance of imputation accuracy, we have restricted our attention to a simple two-population demographic model and a sample size of three sequences. We have also made simplifying assumptions that (1) the two populations and their ancestral population have equal effective population sizes N_e, (2) no migration occurs more recently than the population-divergence time, and (3) no mutation occurs at a nucleotide that has previously experienced a mutation (i.e., the infinite-sites mutation model). Furthermore, we have assumed a straightforward imputation scheme that copies and pastes an entire genomic region of interest in a template reference sequence into the corresponding positions in the target sequence,

rather than allowing different templates in different genomic regions. Each of these assumptions is unlikely to be completely realistic for potential studies in human populations. Nevertheless, the simplicity of our modeling framework has enabled us to analytically study patterns of imputation accuracy that provide insights into the ways in which individual population-genetic factors influence imputation accuracy. These insights, along with continuing development of coalescent-based models for studying genotype imputation, can lead to further insights regarding the performance of imputation methods, and eventually, to advanced strategies for the design of imputation-based association studies in humans and other organisms.

	Branch		
Topology	1	2	3
A	$t_d + t_3$	$t_d + t_3 + 2t_2$	$t_d + t_3$
B	$t_d + t_3 + 2t_2$	$t_d + t_3$	$t_d + t_3$
C	$t_d + t_3$	$t_d + t_3$	$t_d + t_3 + 2t_2$
D	t_3	$2t_d - t_3 + 2t_2$	t_3

Table 5.1: Branch lengths $h_i(\mathcal{T}; g)$ (in units of $2N_e$ generations) for genealogical types $g = A, B, C, D$ and branches $i = 1, 2, 3$ under the two-population model illustrated in Figure 5.1.

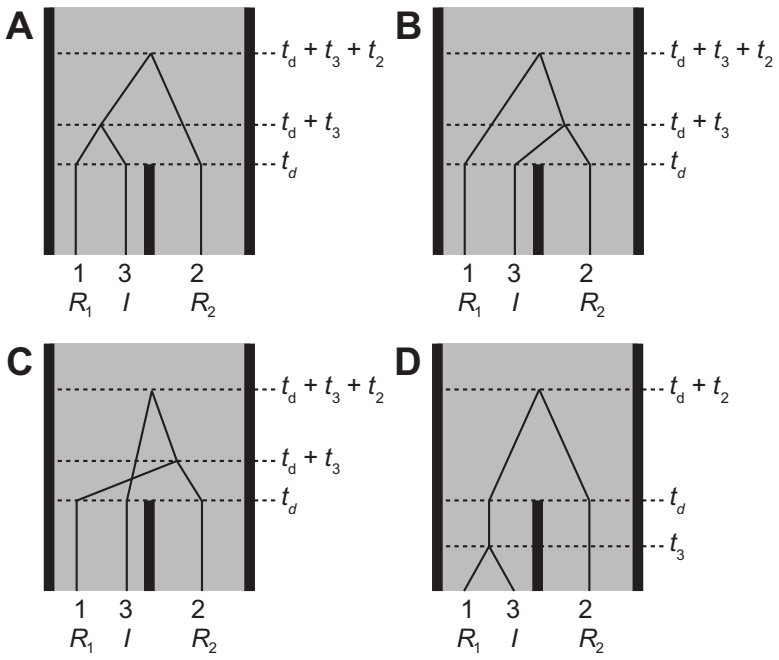

Figure 5.1: Four possible genealogical types for a set of three haploid individuals (a candidate reference individual R_1 and an individual I targeted for imputation from one population, and another candidate reference individual R_2 from a second population). The two populations diverged from an ancestral population at time t_d in the past, and t_k $(k = 3, 2)$ is the length of time during which k distinct lineages exist. Note that in the genealogical types A, B and C, t_3 counts from t_d back in time, whereas in the genealogical type D, T_3 counts from the present.

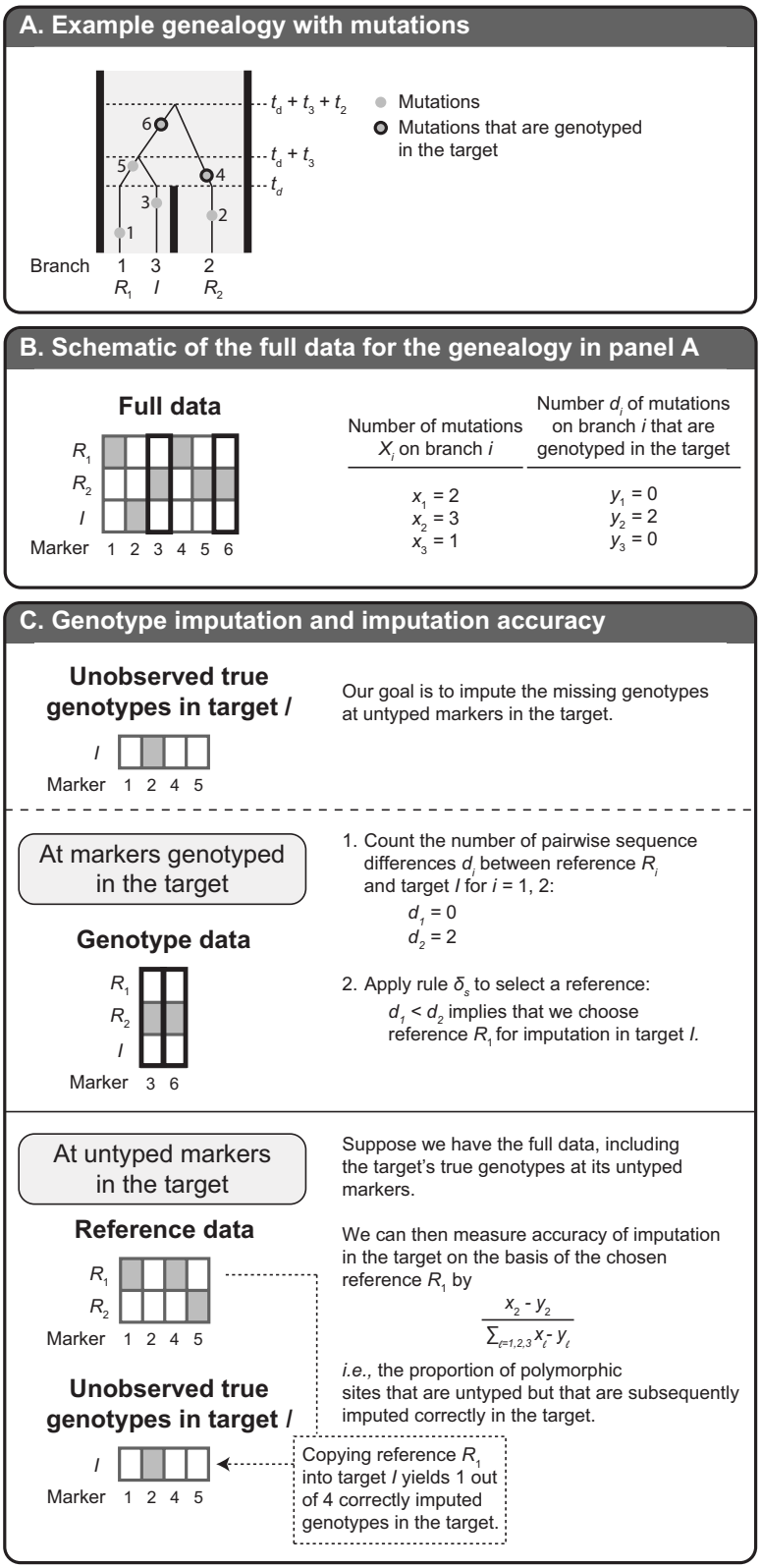

Figure 5.2: Schematic of our imputation procedure. (A) An example genealogy with mutations. (B) A schematic of the full data for the genealogy in (A). (C) An illustration of our imputation procedure. In (B) and (C), white and grey colors indicate ancestral and derived alleles, respectively, and thick black lines indicate genotyped positions in the target.

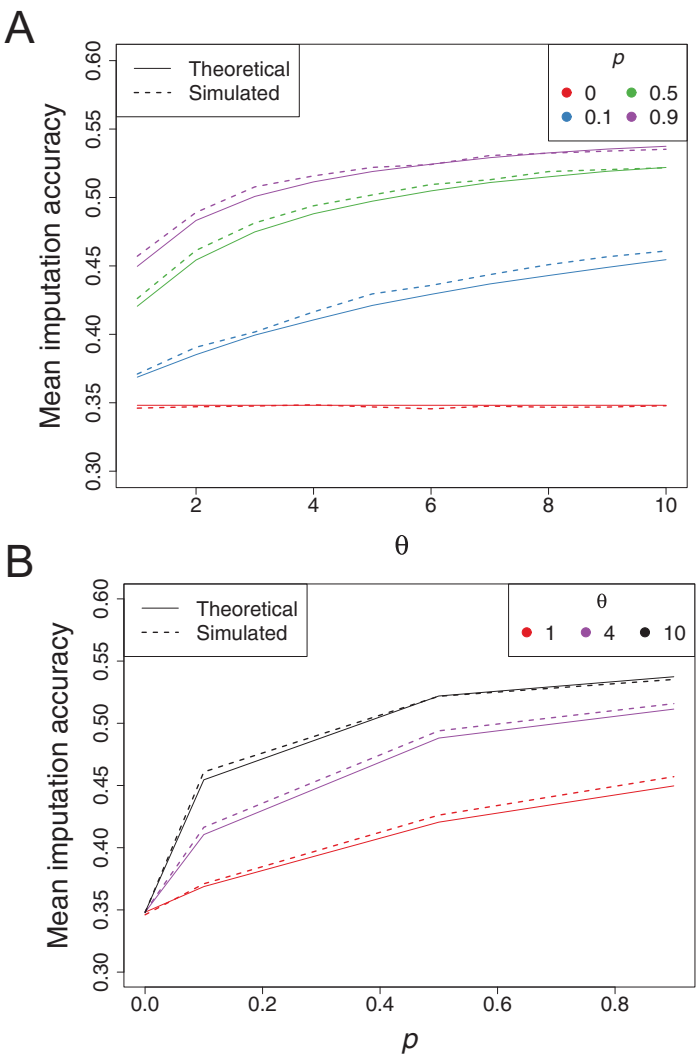

Figure 5.3: The expectated imputation accuracy for various values of a mutation parameter θ and a proportion p of genotypes that are genotyped in the target individual. For $t_d = 0.1$, we obtained theoretical estimates of $\mathbb{E}[Z|\theta, p, t_d]$ using eq. 5.12 (solid line). For $N = 10^5$ and $t_d = 0.1$, we simulated genealogies $\mathcal{G}$ using the algorithm in Section 5.3 to obtain $\bar{z}$, which correspond to Monte Carlo estimates of $\mathbb{E}[Z|\theta, p, t_d]$ (dashed line). We then separately plotted $\mathbb{E}[\widehat{Z|\theta, p, t_d}]$ and $\bar{z}$ as functions of θ and p. The values of θ and p considered were $\{\theta : 1, 2, \ldots, 10\}$ and $\{p : 0, 0.1, 0.5, 0.9\}$. For visual clarity, results are displayed for only selected values of θ when plotted as a function of p. For both plots in this figure, results were obtained from a shared set of simulations.

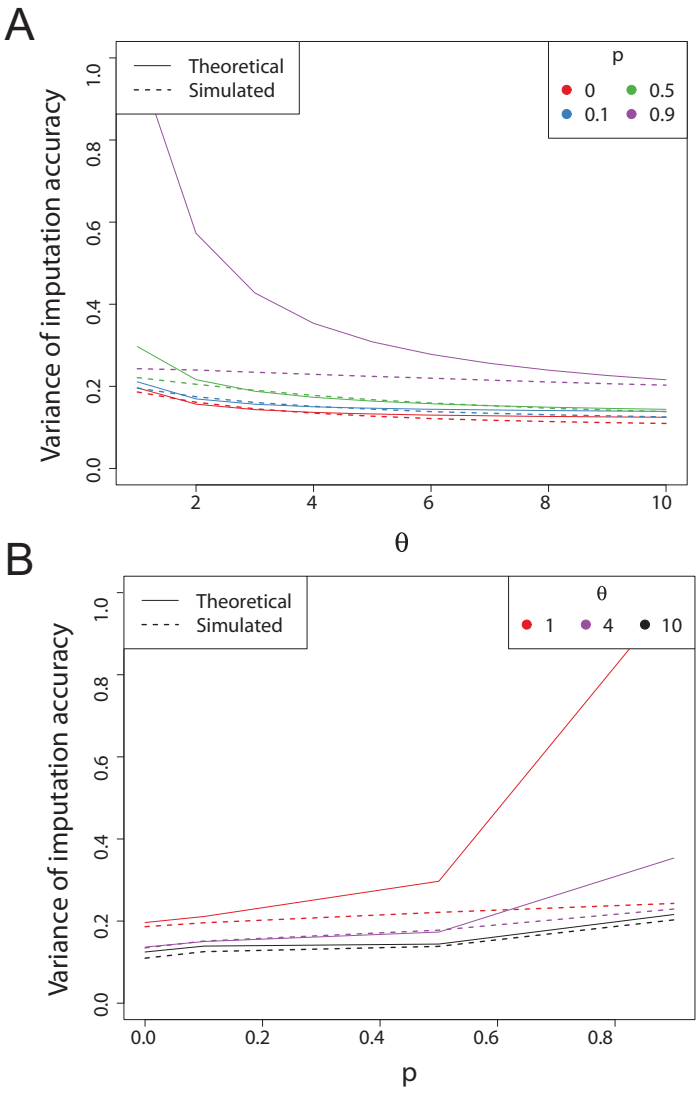

Figure 5.4: The variance of imputation accuracy for various values of θ and p. For $t_d = 0.1$, we obtained theoretical estimates of $\text{Var}[Z|\theta, p, t_d]$ using eq. 5.25 (solid line). For $N = 10^5$ and t_d=0.1, we simulated using the algorithm in Section 5.3 to obtain s^2, which correspond to Monte Carlo estimates of $\text{Var}[Z|\theta, p, t_d]$ (dashed line). We then plotted $\widehat{\text{Var}[Z|\theta, p, t_d]}$ and s^2 as functions of θ and p, separately. The values of θ and p considered were $\{\theta : 1, 2, \ldots, 10\}$ and $\{p : 0, 0.1, 0.5, 0.9\}$. For visual clarity, results are displayed for only selected values of θ when plotted as a function of p. Results of both plots in this figure were obtained from the same simulation as that used to obtain Figure 5.3.

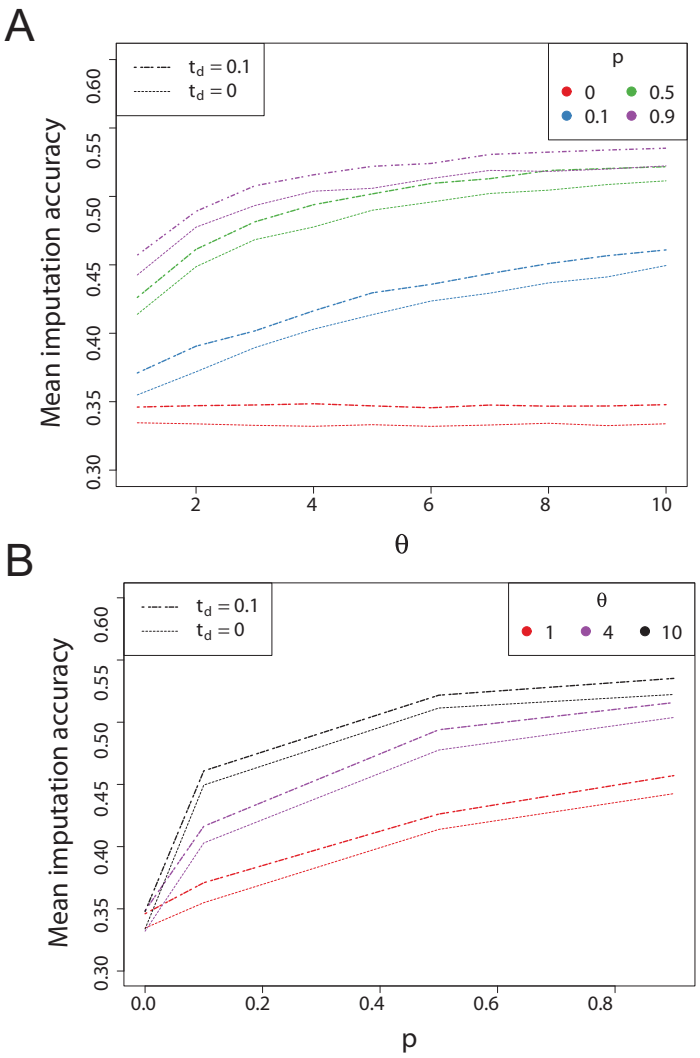

Figure 5.5: The expected imputation accuracy plotted separately (A) as a function of θ and (B) as a function of p for population-divergence times $t_d = 0.1$ and $t_d = 0$. In both plots, results for $t_d = 0.1$ (solid line) and $t_d = 0$ (dashed line) were obtained from the same simulation, with the former values of $\bar{z}$ taken directly from Figure 5.3.

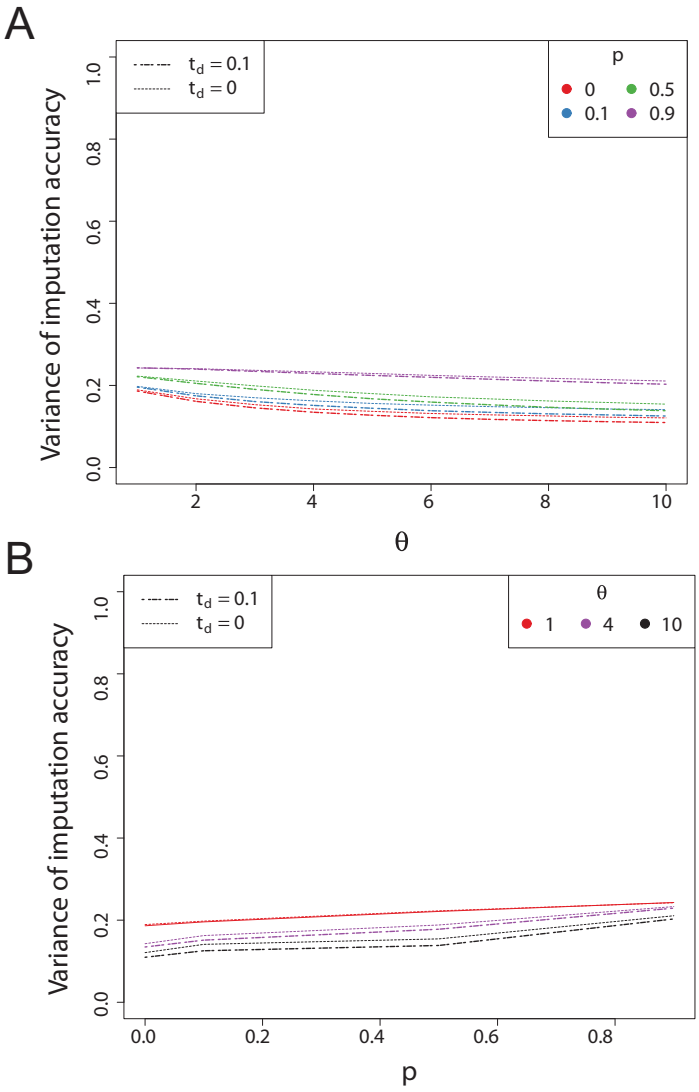

Figure 5.6: The variance of imputation accuracy plotted separately (A) as a function of θ and (B) as a function of p for $t_d = 0.1$ and $t_d = 0$. In both plots, results for $t_d = 0.1$ (solid line) and $t_d = 0$ (dashed line) were obtained from the same simulation, with the former values of s^2 taken directly from Figure 5.4.

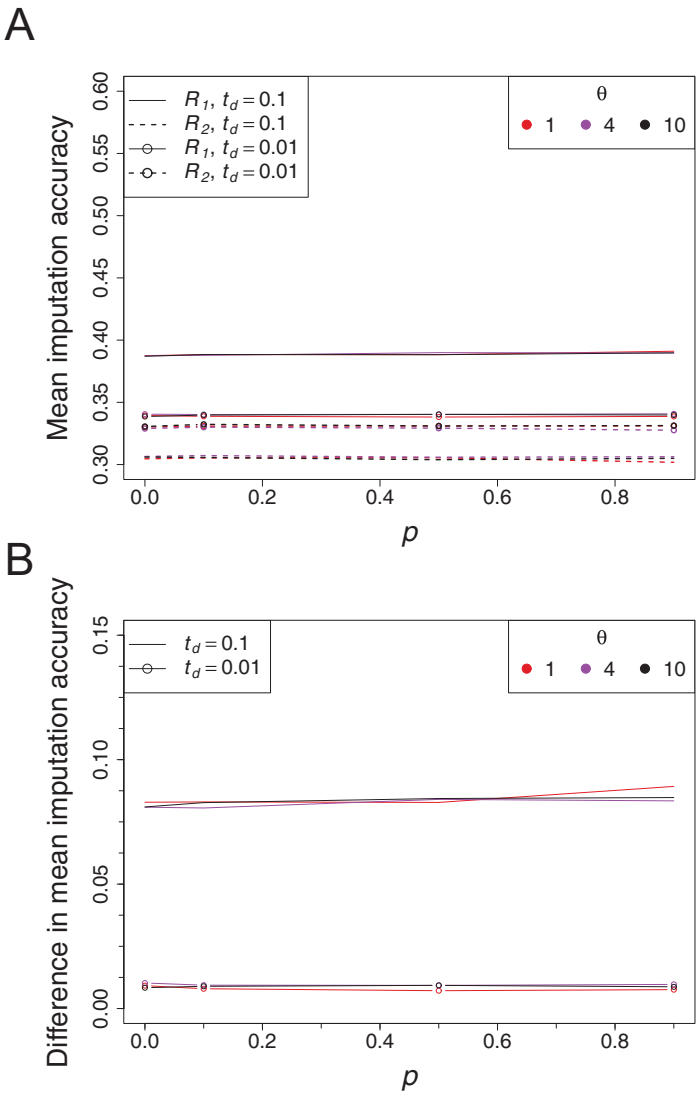

Figure 5.7: The expected accuracy for the imputations separately performed using R_1 and R_2 as the reference sequence, as well as their pointwise differences, plotted as a function of p for $t_d = 0.1$ and $t_d = 0.01$. In both plots, results for $t_d = 0.1$ and $t_d = 0.01$ were obtained from the same simulations used to obtain Figure 5.3.

CHAPTER VI

Conclusion

Genome-wide association (GWA) studies and genotype imputation are two important approaches for elucidating association between genetic variants and complex human diseases. GWA studies, in which several hundred thousand to more than a million genetic markers are assayed in hundreds or thousands of individuals, represent a powerful tool for investigating genetic factors that individually make only small contributions to disease risk. Genotype imputation further allows the evaluation of disease associations at marker positions beyond those measured in individual GWA studies by leveraging information in reference databases of dense genomic data. The success of imputation-based GWA studies has been demonstrated for European populations, in which most discoveries of genes influencing complex phenotypes have occurred. In this dissertation, to extend GWA efforts outside European populations, I studied the design of reference datasets for use in imputation-based genetic association studies that seek to identify disease-predisposing genes in human populations worldwide.

In Chapter II, using dense genotype data from 29 worldwide human populations, I assessed imputation performance in diverse populations and devised an imputation strategy for populations that are poorly represented in existing reference datasets. I showed that the novel imputation strategy—inspired by the unique ancestral histo-

ries of individual populations—increases imputation accuracy in all 29 populations, thereby increasing the potential of GWA studies for uncovering disease-associated genes in these groups. However, I also found that African populations have lower imputation accuracy, suggesting that genomic resources must be expanded for studies in these populations.

In Chapter III, to assess the consequences of imputation error, considering a 2×3 chi-squared test of association, I related imputation error rates to statistical power across the same collection of 29 human populations examined in Chapter II. Unexpectedly, I found that each 1% rise in the error rate requires a substantial increase in the minimal sample size (5-13%) to maintain power. This result suggests that the continuing development of statistical and genomic resources that reduce imputation error will likely translate into substantially reduced sample sizes needed for detecting risk-modifying genes in imputation-based studies of complex diseases.

In Chapter IV, focusing on African populations, I investigated haplotype variation and imputation in Africa, using 253 individuals from 15 Sub-Saharan African populations. Using various statistics on haplotype variation in Sub-Saharan African populations to explain genotype-imputation accuracy observed in the same populations, I found the statistics that measure genetic distance between a target population and candidate reference populations, such as F_{st}, to be useful metrics for guiding the selection of appropriate reference panels for imputation in the target population.

In Chapter V, to analytically study properties of genotype imputation, I developed a coalescent model for evaluating imputation accuracy in terms of population-genetic and study-design parameters. Using this model and a straightforward imputation scheme, I derived the expectation and variance of imputation accuracy, conditioning on a mutation parameter, a proportion of markers genotyped in a given length of a study chromosome, and a time to divergence between study and reference populations. Consistently with prior expectations, the model predicts that on average,

imputation accuracy increases with increasing information for determining, among candidate reference sequences, the reference sequence that is genetically closest to a study sequence targeted for imputation. Interestingly, the model also predicts diminishing returns in improving imputation accuracy from increasing the proportion of markers that are genotyped in a chromosomal region of the study sequence. These results can inform the design of imputation-based association studies and the expansion of public genomic databases.

In summary, this dissertation concerns the development of optimal genotype-imputation strategies for the analysis of large-scale genetic association studies in diverse human populations. My empirical and theoretical findings have the potential to considerably improve the design of genetic studies in populations that are poorly represented in existing public resources. Indeed, results and insights derived from Chapters II and III have aided the identification of multiple novel genetic variants in the gene TMPRSS6 that affect hemoglobin levels in individuals of Indian ancestry (Chambers *et al.*, 2009), as well as genetic variants that contribute to the risk of schizophrenia in subjects of African American ancestry (Shi *et al.*, 2009). In both the study of Chambers *et al.* (2009) and that of Shi *et al.* (2009), consistent with our recommendations, two or more HapMap populations were pooled together to form the appropriate reference panels. Additionally, these chapters have provided arguments for the development of public genomic databases in Mexican and Southeast Asian populations (Silva-Zolezzi *et al.*, 2009; Teo *et al.*, 2009). Chapters IV and V have explored the empirical and theoretical basis for the ways in which population-genetic and study-design parameters influence imputation accuracy, further strengthening the foundation for extending imputation-based association study techniques to populations that have not yet been extensively examined. It is greatly hoped that results from this dissertation will continue to facilitate the search for genetic determinants that influence disease risk in humans.

BIBLIOGRAPHY

BIBLIOGRAPHY

Adeyemo, A., Gerry, N., Chen, G., Herbert, A., Doumatey, A., Huang, H., Zhou, J., Lashley, K., Chen, Y., Christman, M., and Rotimi, C. 2009. A genome-wide association study of hypertension and blood pressure in African Americans, *PLoS Genet.* **5**, e1000564.

Adeyemo, A. and Rotimi, C. 2010. Genetic variants associated with complex human diseases show wide variation across multiple populations, *Public Health Genomics* **13**, 72–79.

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A. M., Hu, T. T., Jiang, R., Muliyati, N. W., Zhang, X., Amer, M. A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J. R., Faure, N., Kniskern, J. M., Jones, J. D. G., Michael, T., Nemri, A., Roux, F., Salt, D. E., Tang, C., Todesco, M., Traw, M. B., Weigel, D., Marjoram, P., Borevitz, J. O., Bergelson, J., and Nordborg, M. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines, *Nature* **465**, 627–631.

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Rotter, J. I., Schumm, L. P., Steinhart, A. H., Targan, S. R., Xavier, R. J., NIDDK IBD Genetics Consortium, Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J.-P., de Vos, M., Vermeire, S., Louis, E., Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Cardon, L. R., Anderson, C. A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N. I., Onnie, C. M., Fisher, S. A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C. G., Parkes, M., Georges, M., and Daly, M. J. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease, *Nat. Genet.* **40**, 955–962.

Becker, T., Flaquer, A., Brockschmidt, F. F., Herold, C., and Steffens, M. 2009. Evaluation of potential power gain with imputed genotypes in genome-wide association studies, *Hum. Hered.* **68**, 23–34.

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., and Cavalli-Sforza, L. L. 1994. High resolution of human evolutionary trees with polymorphic microsatellites, *Nature* **368**, 455–457.

Browning, B. L. and Browning, S. R. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals, *Am. J. Hum. Genet.* **84**, 210–223.

Browning, S. R. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies, *Hum. Genet.* **124**, 439–450.

Browning, S. R. and Browning, B. L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering, *Am. J. Hum. Genet.* **81**, 1084–1097.

Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., Froment, A., Bodo, J.-M., Wambeb, C., Tishkoff, S. A., and Bustamante, C. D. 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans, *Proc Natl Acad Sci USA* **107**, 786–791.

Bustamante, C. D., De La Vega, F. M., and Burchard, E. G. 2011. Genomics for the world, *Nature* **475**, 163–165.

Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G. B., Friedlander, J. S., Groot, H., Gurwitz, D., Jenkins, T., Herrera, R. J., Huang, X., Kidd, J., Kidd, K. K., Langaney, A., Lin, A. A., Mehdi, S. Q., Parham, P., Piazza, A., Pistillo, M. P., Qian, Y., Shu, Q., Xu, J., Zhu, S., Weber, J. L., Greely, H. T., Feldman, M. W., Thomas, G., Dausset, J., and Cavalli-Sforza, L. L. 2002. A human genome diversity cell line panel, *Science* **296**, 261–262.

Casella, G. and Berger, R. L. 2001. "Statistical Inference", Duxbury Press.

Chambers, J. C., Zhang, W., Li, Y., Sehmi, J., Wass, M. N., Zabaneh, D., Hoggart, C., Bayele, H., McCarthy, M. I., Peltonen, L., Freimer, N. B., Srai, S. K., Maxwell, P. H., Sternberg, M. J. E., Ruokonen, A., Abecasis, G., Jarvelin, M.-R., Scott, J., Elliott, P., and Kooner, J. S. 2009. Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels, *Nat. Genet.* **41**, 1170–1172.

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., and Pritchard, J. K. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome, *Nat. Genet.* **38**, 1251–1260.

Cooper, R. S., Tayo, B., and Zhu, X. 2008. Genome-wide association studies: implications for multiethnic samples, *Hum. Mol. Genet.* **17**, R151–R155.

de Bakker, P. I. W., Ferreira, M. A. R., Jia, X., Neale, B. M., Raychaudhuri, S., and Voight, B. F. 2008. Practical aspects of imputation-driven meta-analysis of genome-wide association studies, *Hum. Mol. Genet.* **17**, R122–R128.

Donnelly, P. 2008. Progress and challenges in genome-wide association studies in humans, *Nature* **456**, 728–731.

Druet, T., Schrooten, C., and de Roos, A. P. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle, *J. Dairy Sci.* **93**, 5443–5454.

Egyud, M. R. L., Gajdos, Z. K. Z., Butler, J. L., Tischfield, S., Le Marchand, L., Kolonel, L. N., Haiman, C. A., Henderson, B. E., and Hirschhorn, J. N. 2009. Use of weighted reference panels based on empirical estimates of ancestry for capturing untyped variation, *Hum. Genet.* **125**, 295–303.

Fridley, B. L., Jenkins, G., Deyo-Svendsen, M. E., Hebbring, S., and Freimuth, R. 2010. Utilizing genotype imputation for the augmentation of sequence data, *PLoS ONE* **5**, e11018.

González-Neira, A., Ke, X., Lao, O., Calafell, F., Navarro, A., Comas, D., Cann, H., Bumpstead, S., Ghori, J., Hunt, S., Deloukas, P., Dunham, I., Cardon, L. R., and Bertranpetit, J. 2006. The portability of tagSNPs across populations: a worldwide survey, *Genome Res.* **16**, 323–330.

Gordon, D., Finch, S. J., Nothnagel, M., and Ott, J. 2002. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms, *Hum. Hered.* **54**, 22–33.

Gu, C. C., Yu, K., Ketkar, S., Templeton, A. R., and Rao, D. C. 2008. On transferability of genome-wide tagSNPs, *Genet. Epidemiol.* **32**, 89–97.

Gu, S., Pakstis, A. J., Li, H., Speed, W. C., Kidd, J. R., and Kidd, K. K. 2007. Significant variation in haplotype block structure but conservation in tagSNP patterns among global populations, *Eur. J. Hum. Genet.* **15**, 302–312.

Guan, Y. and Stephens, M. 2008. Practical issues in imputation-based association mapping, *PLoS Genet.* **4**, e1000279.

Hardy, J. and Singleton, A. 2009. Genomewide association studies and human disease, *N. Engl. J. Med.* **360**, 1759–1768.

Henn, B. M., Gignoux, C. R., Jobin, M., Granka, J. M., Macpherson, J. M., Kidd, J. M., Rodriguez-Botigué, L., Ramachandran, S., Hon, L., Brisbin, A., Lin, A. A., Underhill, P. A., Comas, D., Kidd, K. K., Norman, P. J., Parham, P., Bustamante, C. D., Mountain, J. L., and Feldman, M. W. 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans, *Proc. Natl. Acad. Sci. USA* **108**, 5154–5162.

Hindorff, L. A., Junkins, H. A., Mehta, J. P., and Manolio, T. A. Accessed September 6, 2011. A catalog of published genome-wide association studies, *www.genome.gov/gwastudies/* .

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367.

Howie, B. N., Donnelly, P., and Marchini, J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, *PLoS Genet.* **5**, e1000529.

Huang, L., Jakobsson, M., Pemberton, T. J., Ibrahim, M., Nyambo, T., Omar, S., Pritchard, J. K., Tishkoff, S. A., and Rosenberg, N. A. 2012. Haplotype variation and genotype imputation in African populations, *Genet. Epidemiol. (in press)* **00**, 00–00.

Huang, L., Li, Y., Singleton, A. B., Hardy, J. A., Abecasis, G., Rosenberg, N. A., and Scheet, P. 2009a. Genotype-imputation accuracy across worldwide human populations, *Am. J. Hum. Genet.* **84**, 235–250.

Huang, L., Wang, C., and Rosenberg, N. A. 2009b. The relationship between imputation error and statistical power in genetic association studies in diverse populations, *Am. J. Hum. Genet.* **85**, 692–698.

International HapMap 3 Consortium 2010. Integrating common and rare genetic variation in diverse human populations, *Nature* **467**, 52–58.

International HapMap Consortium 2005. A haplotype map of the human genome, *Nature* **437**, 1299–1320.

International HapMap Consortium 2007. A second generation human haplotype map of over 3.1 million SNPs, *Nature* **449**, 851–861.

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H. C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A., and Singleton, A. 2008. Genotype, haplotype and copy-number variation in worldwide human populations, *Nature* **451**, 998–1003.

Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., Kivinen, K., Bojang, K. A., Conway, D. J., Pinder, M., Sirugo, G., Sisay-Joof, F., Usen, S., Auburn, S., Bumpstead, S. J., Campino, S., Coffey, A., Dunham, A., Fry, A. E., Green, A., Gwilliam, R., Hunt, S. E., Inouye, M., Jeffreys, A. E., Mendy, A., Palotie, A., Potter, S., Ragoussis, J., Rogers, J., Rowlands, K., Somaskantharajah, E., Whittaker, P., Widden, C., Donnelly, P., Howie, B., Marchini, J.,

Morris, A., SanJoaquin, M., Achidi, E. A., Agbenyega, T., Allen, A., Amodu, O., Corran, P., Djimde, A., Dolo, A., Doumbo, O. K., Drakeley, C., Dunstan, S., Evans, J., Farrar, J., Fernando, D., Hien, T. T., Horstmann, R. D., Ibrahim, M., Karunaweera, N., Kokwaro, G., Koram, K. A., Lemnge, M., Makani, J., Marsh, K., Michon, P., Modiano, D., Molyneux, M. E., Mueller, I., Parker, M., Peshu, N., Plowe, C. V., Puijalon, O., Reeder, J., Reyburn, H., Riley, E. M., Sakuntabhai, A., Singhasivanon, P., Sirima, S., Tall, A., Taylor, T. E., Thera, M., Troye-Blomberg, M., Williams, T. N., Wilson, M., Kwiatkowski, D. P., Wellcome Trust Case Control Consortium, and Malaria Genomic Epidemiology Network 2009. Genome-wide and fine-resolution association analysis of malaria in West Africa, *Nat. Genet.* **41**, 657–665.

Jewett, E., Zawistowski, M., Rosenberg, N. A., and Zöllner, S. 2012. A coalescent model for imputation reference panel selection, *In preparation* **00**, 00–00.

Johnson, N. L. and Kotz, S. 1969. "Discrete Distributions", Wiley, New York.

Kalinowski, S. T. 2004. Counting alleles with rarefaction: private alleles and hierarchical sampling designs, *Conserv. Genet.* **5**, 539–543.

Kang, S. J., Gordon, D., and Finch, S. J. 2004. What SNP genotyping errors are most costly for genetic association studies?, *Genet. Epidemiol.* **26**, 132–141.

Kirby, A., Kang, H. M., Wade, C. M., Cotsapas, C., Kostem, E., Han, B., Furlotte, N., Kang, E. Y., Rivas, M., Bogue, M. A., Frazer, K. A., Johnson, F. M., Beilharz, E. J., Cox, D. R., Eskin, E., and Daly, M. J. 2010. Fine mapping in 94 inbred mouse strains using a high-density haplotype resource, *Genetics* **185**, 1081–1095.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., San Giovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. 2005. Complement factor H polymorphism in age-related macular degeneration, *Science* **308**, 385–389.

Klionsky, D. J. 2009. Crohns disease, autophagy, and the paneth cell, *N. Engl. J. Med.* **360**, 1785–1786.

Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S., Eyheramendy, S., Voight, B. F., Butler, J. L., Guiducci, C., Illig, T., Hackett, R., Heid, I. M., Jacobs, K. B., Lyssenko, V., Uda, M., The Diabetes Genetics Initiative, FUSION, KORA, The Prostate, Lung Colorectal and Ovarian Cancer Screening Trial, The Nurses' Health Study, SardiNIA, Boehnke, M., Chanock, S. J., Groop, L. C., Hu, F. B., Isomaa, B., Kraft, P., Peltonen, L., Salomaa, V., Schlessinger, D., Hunter, D. J., Hayes, R. B., Abecasis, G. R., Wichmann, H.-E., Mohlke, K. L., and Hirschhorn, J. N. 2008. Identification of ten loci associated with height highlights new biological pathways in human growth, *Nat. Genet.* **40**, 584–591.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., and Myers,

R. M. 2008. Worldwide human relationships inferred from genome-wide patterns of variation, *Science* **319**, 1100–1104.

Li, N. and Stephens, M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data, *Genetics* **165**, 2213–2233.

Li, Y., Ding, J., and Abecasis, G. R. 2006. Mach 1.0: rapid haplotype reconstruction and missing genotype inference, *Am. J. Hum. Genet.* **79**, S2290.

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., and Abecasis, G. R. 2011. Low coverage sequencing: implications for the design of complex trait association studies, *Genome Res.* **21**, doi:10.1101/gr.117259.110.

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes, *Genet. Epidemiol.* **34**, 816–834.

Li, Y., Willer, C. J., Sanna, S., and Abecasis, G. R. 2009. Genotype imputation, *Annu. Rev. Genom. Hum. G.* **10**, 387–406.

Lin, D. Y., Hu, Y., and Huang, B. E. 2008. Simple and efficient analysis of disease association with missing genotype data, *Am. J. Hum. Genet.* **82**, 444–452.

Loos, R. J. F., Lindgren, C. M., Li, S., Wheeler, E., Zhao, J. H., Prokopenko, I., Inouye, M., Freathy, R. M., Attwood, A. P., Beckmann, J. S., Berndt, S. I., Bergmann, S., Bennett, A. J., Bingham, S. A., Bochud, M., Brown, M., Cauchi, S., Connell, J. M., Cooper, C., Smith, G. D., Day, I., Dina, C., De, S., Dermitzakis, E. T., Doney, A. S. F., Elliott, K. S., Elliott, P., Evans, D. M., Sadaf Farooqi, I., Froguel, P., Ghori, J., Groves, C. J., Gwilliam, R., Hadley, D., Hall, A. S., Hattersley, A. T., Hebebrand, J., Heid, I. M., Herrera, B., Hinney, A., Hunt, S. E., Jarvelin, M.-R., Johnson, T., Jolley, J. D. M., Karpe, F., Keniry, A., Khaw, K.-T., Luben, R. N., Mangino, M., Marchini, J., McArdle, W. L., McGinnis, R., Meyre, D., Munroe, P. B., Morris, A. D., Ness, A. R., Neville, M. J., Nica, A. C., Ong, K. K., O'Rahilly, S., Owen, K. R., Palmer, C. N. A., Papadakis, K., Potter, S., Pouta, A., Qi, L., Randall, J. C., Rayner, N. W., Ring, S. M., Sandhu, M. S., Scherag, A., Sims, M. A., Song, K., Soranzo, N., Speliotes, E. K., Syddall, H. E., Teichmann, S. A., Timpson, N. J., Tobias, J. H., Uda, M., Ganz Vogel, C. I., Wallace, C., Waterworth, D. M., Weedon, M. N., Willer, C. J., Wraight, V. L., Yuan, X., Zeggini, E., Hirschhorn, J. N., Strachan, D. P., Ouwehand, W. H., Caulfield, M. J., Samani, N. J., Frayling, T. M., Vollenweider, P., Waeber, G., Mooser, V., Deloukas, P., McCarthy, M. I., Wareham, N. J., and Barroso, I. 2008. Common variants near *MC4R* are associated with fat mass, weight and risk of obesity, *Nat. Genet.* **40**, 768–775.

Manolio, T. A. 2010. Genomewide association studies and assessment of the risk of disease, *N. Engl. J. Med.* **363**, 166–176.

Manolio, T. A., Brooks, L. D., and Collins, F. S. 2008. A HapMap harvest of insights into the genetics of common disease, *J. Clin. Invest.* **118**, 1590–1605.

Marchini, J. and Howie, B. 2010. Genotype imputation for genome-wide association studies, *Nat. Rev. Genet.* **11**, 499–511.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes, *Nat. Genet.* **39**, 906–913.

Montes, T., Tortajada, A., Morgan, B. P., Rodriguez de Cordoba, S., and Harris, C. L. 2009. Functional basis of protection against age-related macular degeneration conferred by a common polymorphism in complement factor b, *Proc. Natl. Acad. Sci. USA* **106**, 4366–4371.

Need, A. C. and Goldstein, D. B. 2009. Next generation disparities in human genomics: concerns and remedies, *Trends Genet.* **25**, 489–494.

Nicolae, D. L. 2006. Testing untyped alleles (TUNA)—applications to genome-wide association studies, *Genet. Epidemiol.* **30**, 718–727.

Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M., and Franke, A. 2009. A comprehensive evaluation of SNP genotype imputation, *Hum. Genet.* **125**, 163–171.

Paşaniuc, B., Avinery, R., Gur, T., Skibola, C., Bracci, P. M., and Halperin, E. 2010. A generic coalescent-based framework for the selection of a reference panel for imputation, *Genet. Epidemiol.* **34**, 773–782.

Pei, Y.-F., Li, J., Zhang, L., Papasian, C. J., and Deng, H.-W. 2008. Analyses and comparison of accuracy of different genotype imputation methods, *PLoS ONE* **3**, e3551.

Pemberton, T. J., Jakobsson, M., Conrad, D. F., Coop, G., Wall, J. D., Pritchard, J. K., Patel, P. I., and Rosenberg, N. A. 2008. Using population mixtures to optimize the utility of genomic databases: linkage disequilibrium and association study design in India, *Ann. Hum. Genet.* **72**, 535–546.

Pemberton, T. J., Wang, C., Li, J. Z., and Rosenberg, N. A. 2010. Inference of unexpected genetic relatedness among individuals in HapMap Phase III, *Am. J. Hum. Genet.* **87**, 457–464.

Pennisi, E. 2007. Breakthrough of the year: human genetic variation, *Science* **318**, 1842–1843.

Pritchard, J. K. and Przeworski, M. 2001. Linkage disequilibrium in humans: models and data, *Am. J. Hum. Genet.* **69**, 1–14.

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., and Cavalli-Sforza, L. L. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa, *Proc. Natl. Acad. Sci. USA* **102**, 15942–15947.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. 2001. Linkage disequilibrium in the human genome, *Nature* **411**, 199–204.

Reiner, A. P., Barber, M. J., Guan, Y., Ridker, P. M., Lange, L. A., Chasman, D. I., Walston, J. D., Cooper, G. M., Jenny, N. S., Rieder, M. J., Durda, J. P., Smith, J. D., Novembre, J., Tracy, R. P., Rotter, J. I., Stephens, M., Nickerson, D. A., and Krauss, R. M. 2008. Polymorphisms of the *HNF1A* gene encoding hepatocyte nuclear factor-1α are associated with C-reactive protein, *Am. J. Hum. Genet.* **82**, 1193–1201.

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., and Boehnke, M. 2010. Genome-wide association studies in diverse populations, *Nat. Rev. Genet.* **11**, 356–366.

Scheet, P. and Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase, *Am. J. Hum. Genet.* **78**, 629–644.

Scheinfeldt, L. B., Soi, S., and Tishkoff, S. A. 2010. Working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history, *Proc. Natl. Acad. Sci. USA* **107**, 8931–8938.

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., Erdos, M. R., Stringham, H. M., Chines, P. S., Jackson, A. U., Prokunina-Olsson, L., Ding, C.-J., Swift, A. J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X.-Y., Conneely, K. N., Riebow, N. L., Sprau, A. G., Tong, M., White, P. P., Hetrick, K. N., Barnhart, M. W., Bark, C. W., Goldstein, J. L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T. A., Watanabe, R. M., Valle, T. T., Kinnunen, L., Abecasis, G. R., Pugh, E. W., Doheny, K. F., Bergman, R. N., Tuomilehto, J., Collins, F. S., and Boehnke, M. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants, *Science* **316**, 1341–1345.

Servin, B. and Stephens, M. 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits, *PLoS Genet.* **3**, e114.

Shi, J., Levinson, D. F., Duan, J., Sanders, A. R., Zheng, Y., Pe'er, I., Dudbridge, F., Holmans, P. A., Whittemore, A. S., Mowry, B. J., Olincy, A., Amin, F., Cloninger, C. R., Silverman, J. M., Buccola, N. G., Byerley, W. F., Black, D. W., Crowe, R. R., Oksenberg, J. R., Mirel, D. B., Kendler, K. S., Freedman, R., and Gejman, P. V. 2009. Common variants on chromosome 6p22.1 are associated with schizophrenia, *Nature* **460**, 753–757.

Shriner, D., Adeyemo, A., Chen, G., and Rotimi, C. N. 2010. Practical considerations for imputation of untyped markers in admixed populations, *Genet. Epidemiol.* **34**, 258–265.

Silva-Zolezzi, I., Hidalgo-Miranda, A., Estrada-Gil, J., Fernandez-Lopez, J. C., Uribe-Figueroa, L., Contreras, A., Balam-Ortiz, E., del Bosque-Plata, L., Velazquez-Fernandez, D., Lara, C., Goya, R., Hernandez-Lemus, E., Davila, C., Barrientos, E., March, S., and Jimenez-Sanchez, G. 2009. Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico, *Proc. Natl. Acad. Sci. USA* **106**, 8611–8616.

Stephens, J. C., Schneider, J. A., Tanguay, D. A., Choi, J., Acharya, T., Stanley, S. E., Jiang, R., Messer, C. J., Chew, A., Han, J.-H., Duan, J., Carr, J. L., Lee, M. S., Koshy, B., Kumar, A. M., Zhang, G., Newell, W. R., Windemuth, A., Xu, C., Kalbfleisch, T. S., Shaner, S. L., Arnold, K., Schulz, V., Drysdale, C. M., Nandabalan, K., Judson, R. S., Ruaño, G., and Vovis, G. F. 2001. Haplotype variation and linkage disequilibrium in 313 human genes, *Science* **293**, 489–493.

Stephens, M. and Scheet, P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation, *Am. J. Hum. Genet.* **76**, 449–462.

Stranger, B. E., Stahl, E. A., and Raj, T. 2011. Progress and promise of genome-wide association studies for human complex trait genetics, *N. Engl. J. Med.* **187**, 367–383.

Surakka, I., Kristiansson, K., Anttila, V., Inouye, M., Barnes, C., Moutsianas, L., Salomaa, V., Daly, M., Palotie, A., Peltonen, L., and Ripatti, S. 2010. Founder population-specific hapmap panel increases power in GWA studies through improved imputation accuracy and CNV tagging, *Genome Res.* **20**, 1344–1351.

Tang, H. 2006. Confronting ethnicity-specific disease risk, *Nat. Genet.* **38**, 13–15.

Teo, Y.-Y., Sim, X., Ong, R. T. H., Tan, A. K. S., Chen, J., Tantoso, E., Small, K. S., Ku, C.-S., Lee, E. J. D., Seielstad, M., and Chia, K.-S. 2009. Singapore Genome Variation Project: A haplotype map of three Southeast Asian populations, *Genome Res.* **19**, 2154–2162.

Teo, Y.-Y., Small, K. S., and Kwiatkowski, D. P. 2010. Methodological challenges of genome-wide association analysis in Africa, *Nat. Rev. Genet.* **11**, 149–160.

The 1000 Genomes Project Consortium 2010. A map of human genome variation from population-scale sequencing, *Nature* **467**, 1061–1073.

Tishkoff, S. A. and Kidd, K. K. 2004. Implications of biogeography of human populations for 'race' and medicine, *Nat. Genet.* **36**, S21–S27.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Boda, J.-M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L., and Williams, S. M. 2009. The genetic structure and history of Africans and African Americans, *Science* **324**, 1035–1044.

Wakeley, J. 2008. "Coalescent Theory", Roberts & Company, Greenwood Village, CO.

Weir, B. S. 1996. "Genetic Data Analysis II", Sinauer Associates, Sunderland, Massachusetts, USA.

Wellcome Trust Case Control Consortium 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature* **447**, 661–678.

Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M., Strait, J., Duren, W. L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A. J., Morken, M. A., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Herçberg, S., Zelenika, D., Chen, W.-M., Li, Y., Scott, L. J., Scheet, P. A., Sundvall, J., Watanabe, R. M., Nagaraja, R., Ebrahim, S., Lawlor, D. A., Ben-Shlomo, Y., Davey-Smith, G., Shuldiner, A. R., Collins, R., Bergman, R. N., Uda, M., Tuomilehto, J., Cao, A., Collins, F. S., Lakatta, E., Lathrop, G. M., Boehnke, M., Schlessinger, D., Mohlke, K. L., and Abecasis, G. R. 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease, *Nat. Genet.* **40**, 161–169.

Xing, J., Witherspoon, D. J., Watkins, W. S., Zhang, Y., Tolpinrud, W., and Jorde, L. B. 2008. HapMap tagSNP transferability in multiple populations: general guidelines, *Genomics* **92**, 41–51.

Yano, T. and Kurata, S. 2009. An unexpected twist for autophagy in crohn's disease, *Nat. Immunol.* **10**, 134–136.

Yu, Z. and Schaid, D. J. 2007. Methods to impute missing genotypes for population data, *Hum. Genet.* **122**, 495–504.

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., and Diabetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes, *Nat. Genet.* **5**, 638–645.

Zhao, Z., Timofeev, N., Hartley, S. W., Chui, D. H. K., Fucharoen, S., Perls, T. T., Steinberg, M. H., Baldwin, C. T., and Sebastiani, P. 2008. Imputation of missing genotypes: an empirical evaluation of IMPUTE, *BMC Genet.* **9**, 85.